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

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 [8]. Lists or matrices of numbers may be able to communicate simple trends, but for more complex patterns, harnessing the human visual system [56] 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 [48], to present evidence in court cases [13], and to facilitate collaboration and encourage engagement on climate change issues [61, 82]. 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 [31]. 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 [13, 24, 26, 44, 45, 55, 86] and estimation via bisection tasks [76], 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* [88] 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* [87] is reproduced in Chapter 5. Sections 5.5, 5.6, and 5.7, contain minimally altered parts of the pub-

lished article.

- *Effects of Point Size and Opacity Adjustments in Scatterplots* [89] is reproduced in Chapter 6. Sections 6.5, **??**, and 6.7 contain minimally altered parts of the published article.
- *Effects of Alternative Scatterplot Designs on Belief* [**strain_2025**] is reproduced in Chapter 7. Sections **??**, 7.5.2, 7.5.3, and 7.5.4, 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

Related Work

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 [22], 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

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

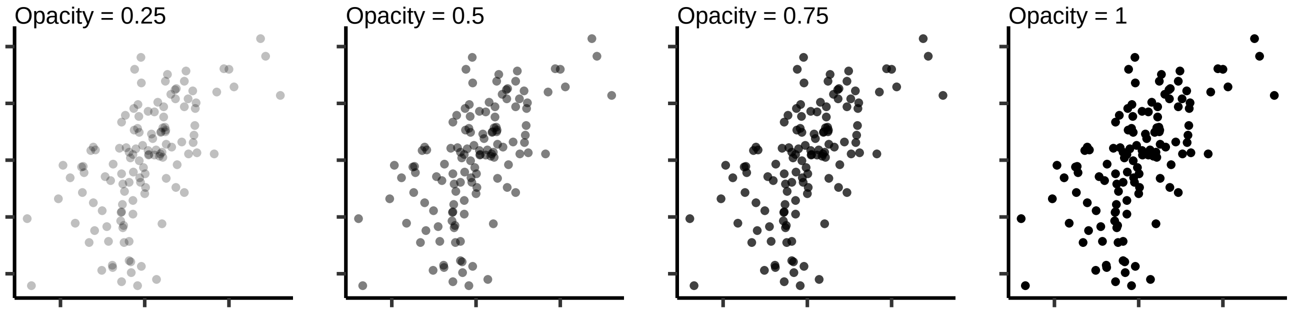

Figure 3.2. Examples of the stimuli used in experiment 1, demonstrated with an r value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.

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 [7], closed-source software, such as Gorilla [5] or E-prime [28] 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 [66] due to its open-source status, flexibility regarding graphical and code-based experimental design, and high level of timings accuracy [16]. 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 researchers.

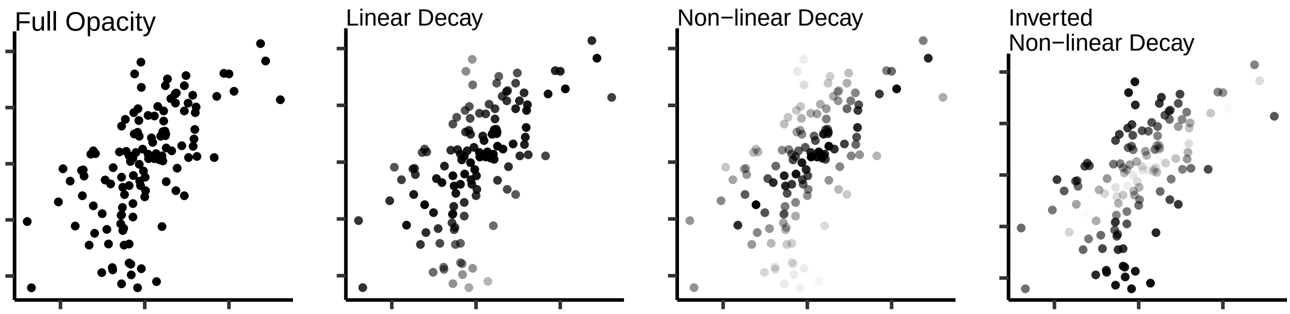

Figure 3.3. Examples of the stimuli used in experiment 2, demonstrated with an r value of 0.6.

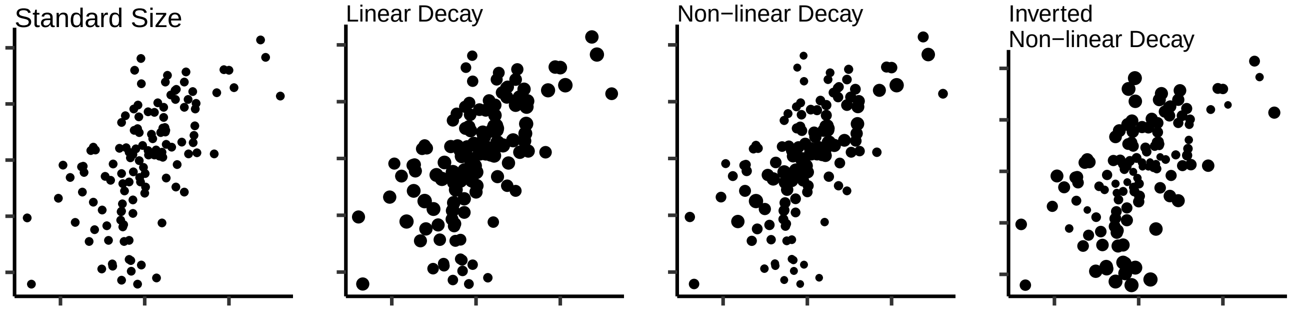

Figure 3.4. Examples of the stimuli used in experiment 3, demonstrated with an *r* value of 0.6.

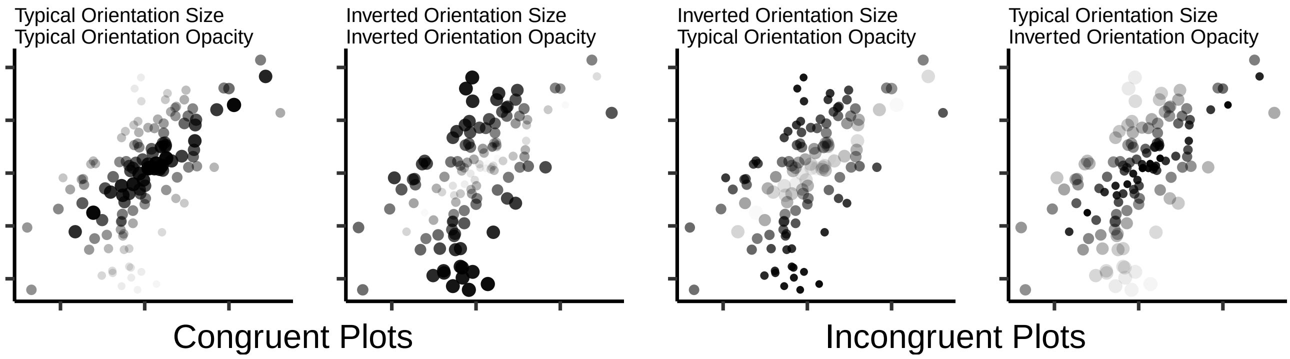

Figure 3.5. Examples of the stimuli used in experiment 4, demonstrated with an *r* value of 0.6.

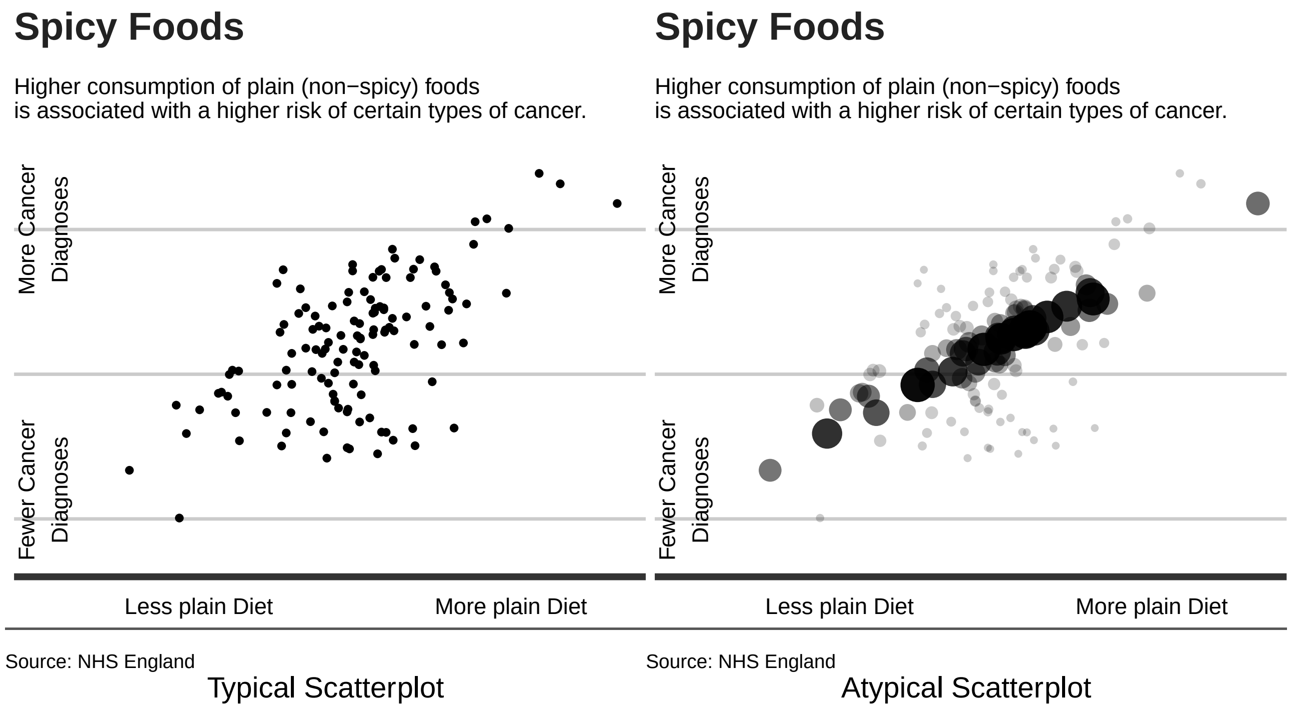

Figure 3.6. Examples of the experimental stimuli for experiment 5 using an *r* value of 0.6. Group A saw the alternative scatterplot presented on the right, while group B saw the typical design on the left.

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 [6, 36, 71], 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 [72] and CloudResearch provide the highest quality data for the lowest cost [27]; 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 [21], 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 [65] 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 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.

[1] this is a more strict rate than the 95% recommended by Peer et al. [65].

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 [18]. 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` [100] in R. Specific versions numbers are provided 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 $r = 0.2$ [13, 24, 86]. 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 discussions and analyses.

3.3 Analytical Methods

All analyses in this thesis were conducted using R (version 4.4.2 [73]). 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 modelling is a reliable approach that is resistant to a variety of distributional assumption violations [81], and facilitates the appreciation of the data story in a broader and more detailed fashion.

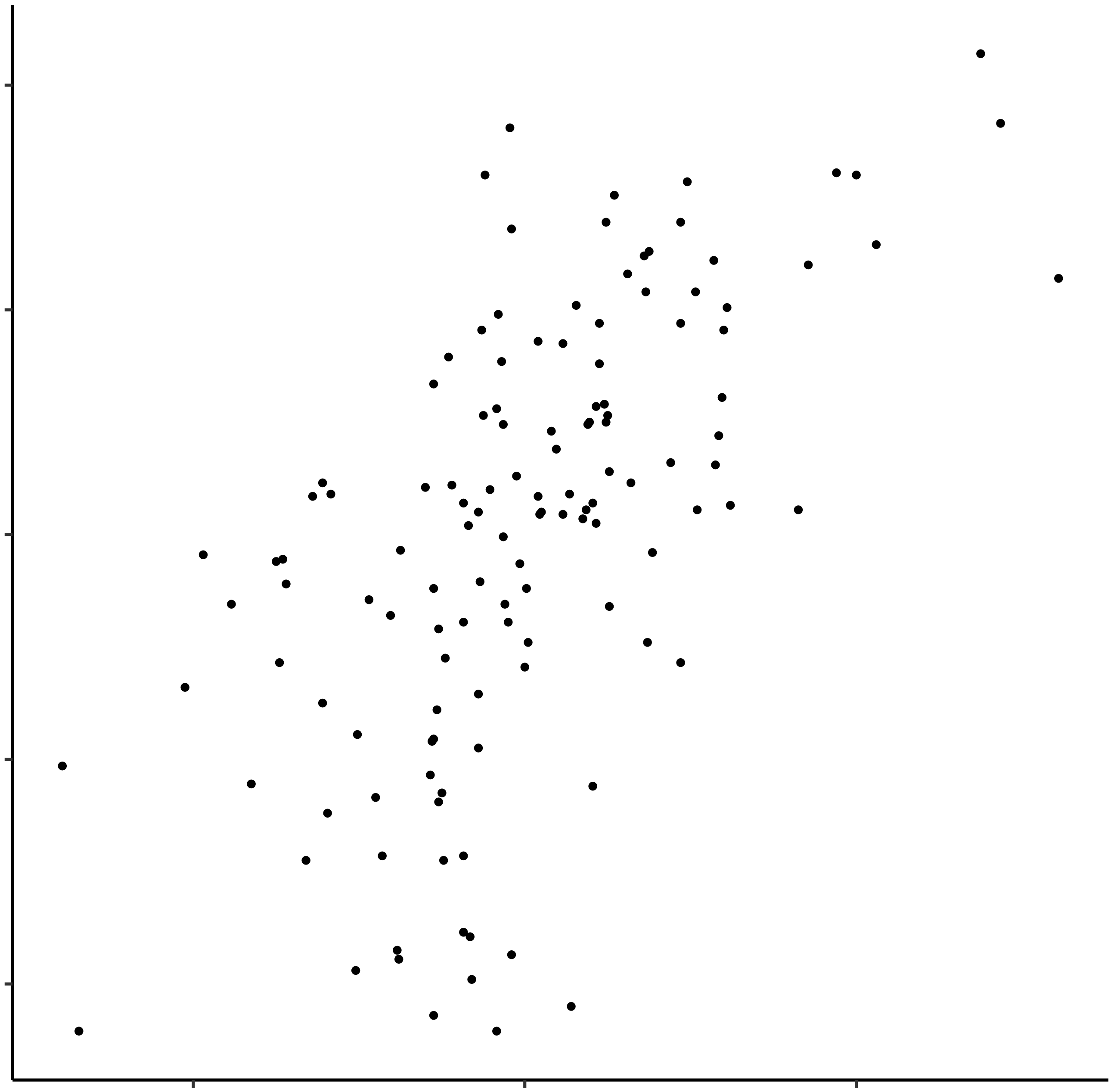

Figure 3.7. The basic design of scatterplots in experiments 1 to 4.

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 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 [17]. Figure 3.8 visualises these concepts.

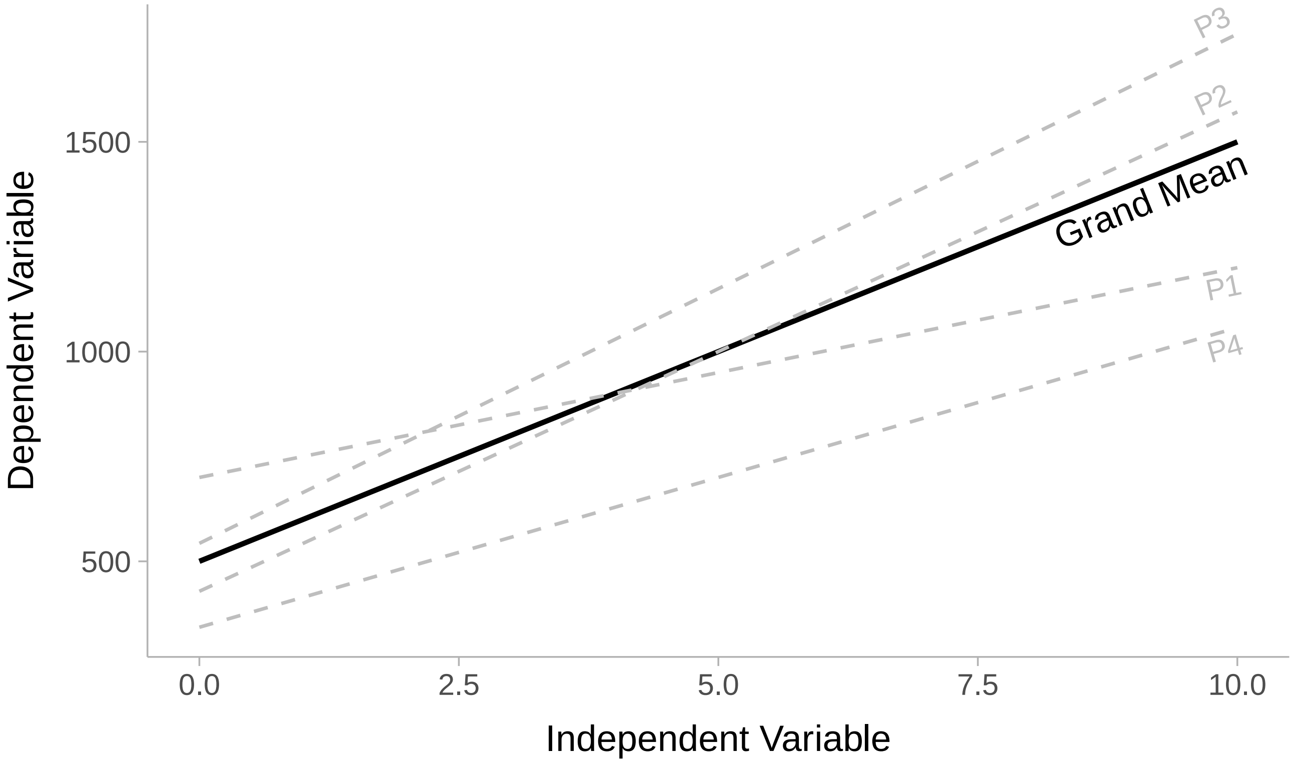

Figure 3.8. Visualising random intercepts and slopes for a theoretical experiment with 4 participants. The grand mean of the dependent variable is shown as a solid line, while each separate random intercept is drawn with dashed lines. Each line has a different gradient, representing different random slopes for each participant. This graphic was inspired by those featured in Brown, 2021 [17].

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 [84]; 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 [47] 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 [49]. In light of these issues, the `ordinal` package [23] 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. [9] 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. [10] 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 [95] 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 [41] 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 [39] 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 [59]. 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 [11]. 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 [54].

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 [43] 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 [69].

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` [46], written in 2002, allowed R code to be integrated into LaTeX documents. This was followed by Yihui Xie's `knitr` [102], which expanded `Sweave` functionality and improved integration with tools such as `pandoc` [51]. `knitr` uses `Rmarkdown` [103] 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` [3], 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 [37]. 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 [25, 79, 93]. 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 [14, 53]. 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 [53]. 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 [15] 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 [68] (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 [62] crystallised in the early 2010s [67], concerns had been voiced since at least the late 1960s [78]. 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 [80]). These issues led this project to strive for a gold standard [68] 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 [4, 42] and secondary use of data [91], and guards against reproducibility issues [57]. Quite aside from external motivating factors, I found that developing and embedding the reproducibility practices described here have resulted

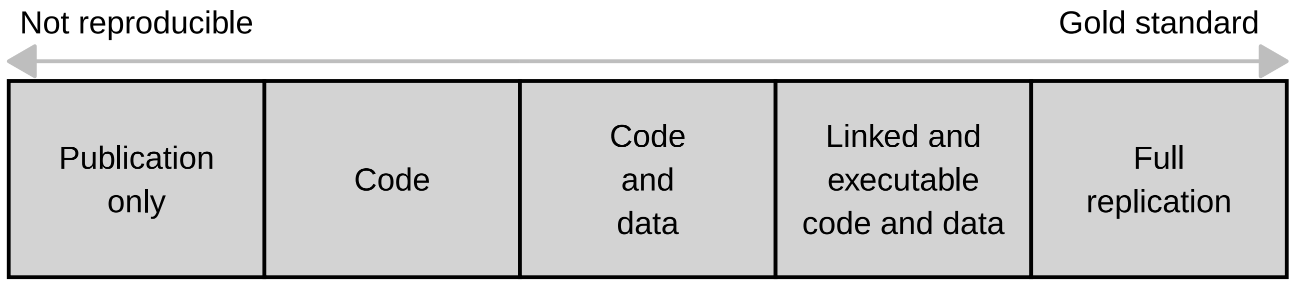

Figure 3.9. Peng's (2011) Reproducibility Spectrum. This figure has been reproduced from Peng (2011) [68].

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 [92] 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 [40].

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 [50], 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 [64]. All hypotheses and analysis plans were pre-registered with the Open Science Framework [63]. 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 [101] 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 sufficiently to 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 proposal is included in the supplemental materials. This was prompted by recent research demonstrating consistent perceptual biases in scatterplots due to geometric scaling [99], the growing 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, however 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

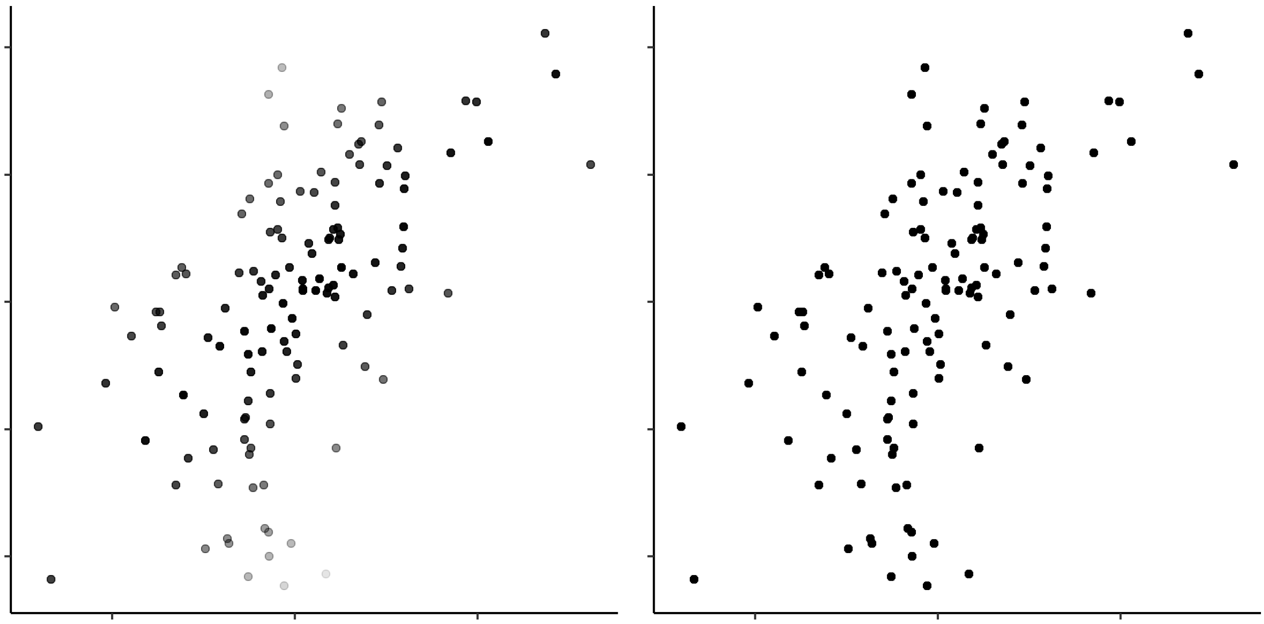

Figure 4.1. Examples of the experimental stimuli used in the pilot study (opacity decay factor). On the left, the opacity decay function is visible. Note the linear scaling used.

on the effects of changing levels of opacity in various parts of a plot on correlation estimation. This 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 were predicted by plot size and the presence or absence 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` [47] 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 of correlation estimation error 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).

Contrast		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

Contrast		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 and between medium and small plots. There were no significant differences in correlation estimation performance between large and medium 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 [99]. 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 I 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 [38], 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 the established effect of plot 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 [74, 75], 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

Findings from the pilot study suggest that changing the opacities of points in scatterplots is able to change participants' estimates of the correlation being displayed. The effect found in that study was too small to make a real difference with regards to correcting for the underestimation bias, and does not provide information on *how* changing opacity might change the percept (only that *it does*). Failing to understand the ways in which opacity is able to change the perception of correlation prevents future

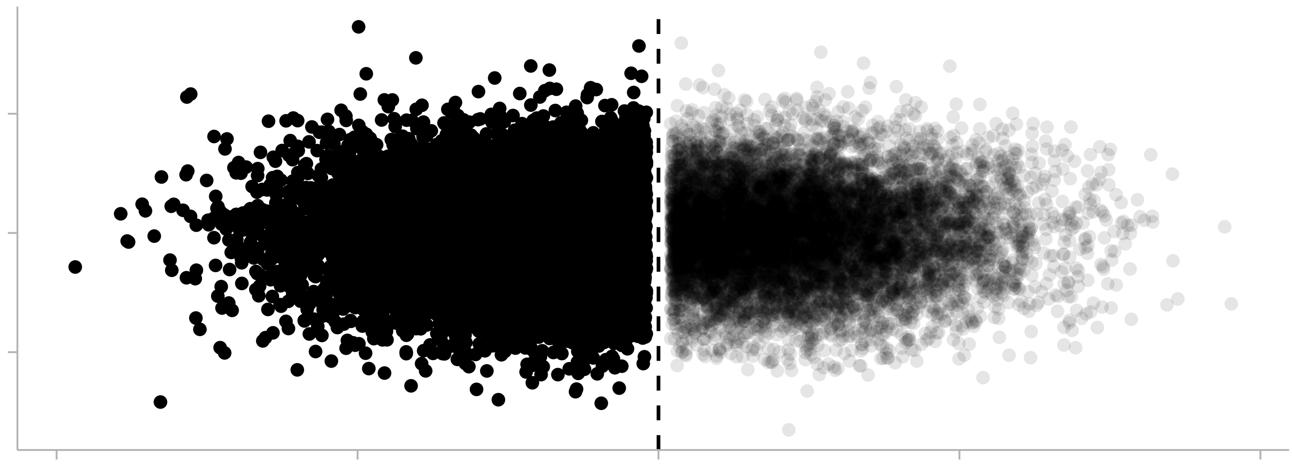

Figure 4.2. Adjusting point opacity to address overplotting. Contrast between the points and the background is full (alpha = 1, full opacity points, Left) or low (alpha = .1, low opacity points, Right). The dataset used has 20,000 points.

work from tuning what was a small effect in the pilot study into something with real potential for producing more perceptually optimised scatterplots.

4.3.1 Overview

In two experiments, the opacities of points in scatterplots were 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 partially correct for a historic underestimation bias.

4.4 Related Work

4.4.1 Transparency, Contrast, Opacity, and Formal Definitions

The original paper that this chapter is based on is titled "The Effects of Contrast on Correlation Perception in Scatterplots". In response to reviewer comments to the paper that forms , the term "contrast" was changed to "opacity". In order to maintain consistency throughout this thesis, the more up-to-date wording (opacity) is used, although I discuss the issue of terminology below.

Adjusting the opacity of points in scatterplots is an established technique used to address issues of overplotting or clutter [12, 52], in which scatterplots with large numbers of data points suffer from visibility issues caused by excessive point density. Lowering the opacity of all scatterplot points using alpha blending [30] addresses this, and makes data trends and distributions easier to see and interpret for the reader. Figure 4.2 demonstrates the impact of lowering global point opacity in a scatterplot with a very high number of data points.

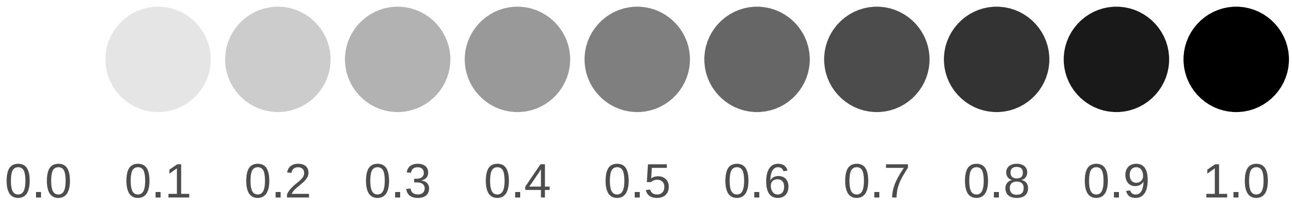

Figure 4.3. The relationship between alpha values and rendered point opacity. Higher alpha values result in greater contrast between the foreground (scatterplot point) and background. When alpha = 0, the foreground is ignored and the background is rendered.

Lowering opacity leads to a reduction of the contrast of isolated points with the background, and for regions with overlapping points, colour intensities are summed. The stimuli used in the experiments throughout this chapter had 128 small points, meaning the majority of points were clearly visible at all times. For this reason, the effects of point overlap were not taken into account when designing and analysing the experiments described here. Due to this, the approach to opacity described in this chapter would not be useful when dealing with much larger datasets where clutter becomes an issue.

The `ggplot2` [100] package (version 3.4.1) was used in R to create stimuli for this experiment. This package uses an alpha parameter to set point opacity. Alpha here refers to the level of linear interpolation [85] between foreground and background pixel values; alpha values of 0 (full transparency) and 1 (full opacity) result in no interpolation and rendering of either the background or foreground pixel values respectively. Alpha values between 0 and 1 correspond to different ratios of interpolation, and are illustrated in Figure 4.3.

Definitions of contrast, opacity, and transparency are fuzzy. Often, different works will use the terms interchangeably. As mentioned above, I initially elected to use the term “contrast”, given the fundamentality of contrast as a feature of human visual perception [33], however later reviewer comments prompted the change to “opacity”. Nevertheless, we can consider *opacity* as it is used here when pertaining to scatterplot points on a white background to be shorthand for “the contrast between foreground and background objects”, as visually, these concepts are the same. There are numerous psychophysical definitions of perceived contrast [105] based on what is being presented, for example, models that take into account visibility limits (CIELAB lightness), or contrast in periodic patterns such as sinusoidal gratings (Michelson’s contrast). The common thread running through these definitions is the use of a ratio between target and background luminances. The experiments described here take place online, with participants completing experiments on their personal laptop or desktop computers. Due to this, the experimenter has no control over the exact luminances of stimuli, only over the relative luminance between targets (scatterplot points) and backgrounds. Given my interest in relative differences in correlation perception averaged over a series of 180 single-plot trials, this lack of control over the exact nature of the stimuli was not problematic. It does mean that reporting the exact luminance values would be pointless however, so where a value for opacity is referred to in this chapter and beyond, it is the alpha value specified by `ggplot2`.

4.4.2 Effects of Point Opacity on Correlation Estimation

Despite the popularity of adjusting opacity to address overplotting issues, little investigation had taken place into the effects of reducing point opacity on people's perceptions of correlation. In 2012, Rensink [74] found correlation perception to be invariant to changes in point opacity, although this work took place with only small sample sizes (n = 12), and using only bisection/JND methodologies (see section in related work about ways of testing correlation perception).

Changing the contrast between a stimulus and its background (lowering its opacity) effectively reduces the strength of its signal. A likely consequence of this is greater levels of uncertainty in aspects of that stimulus, for example, the locations of points in scatterplots. Consequently, one might anticipate that increased uncertainty could lead to altered perceptions of correlation and/or the presence of greater levels of noise in correlation estimates due to effects on the perceived position of points within a scatterplot point cloud. While there is evidence [98] that perception of stimulus position becomes exponentially worse as contrast is reduced (as measured by vernier acuity tasks), this is only true for a narrow range of low contrast stimuli just above the detection threshold. For stimuli that feature higher contrasts between them and their backgrounds, vernier acuity appears robust to such changes. Nevertheless, there is evidence that other perceptual estimates become more uncertain with reduced contrast, such as speed perception [20]. With this in mind, I argue that the effects of stimulus opacity on perceived correlation in scatterplots warrants further investigation.

In 2022, Hong et al. [38], used point opacity and size to encode a third variable in trivariate scatterplots, while asking participants to judge the average position of all the points displayed. It was found that participants' estimates of average point position were biased towards areas of larger or darker points; this was termed the *weighted average illusion*. Together with evidence that darker (more opaque) and larger points are more salient [**healey_2011**], the implication I took from this work was that I could use point opacity to systematically lower the salience of the points representing the widest parts of the probability distribution; if participants promptly perceived a narrower distribution, I expected this to be able to (at least partially) correct for the underestimation bias.

One way to correct for an underestimation of correlation in scatterplots would be to simply remove outer data points until correlation perception is aligned with the actual correlation value. However, this would necessitate hiding data and thus changing the information presented to the viewer. An alternative approach is to manipulate the opacity of only some of the points; it would seem most sensible to do so for the points that are more extreme relative to the underlying regression line. In the present study these questions are explored in two online experiments with large sample sizes. In the first, the effects of point opacity over the entire scatterplot on correlation estimates is investigated. The second experiment examines how changing contrast as a function of distance to the regression line affects perceived correlation. To pre-empt the results, clear effects of both manipulations are found.

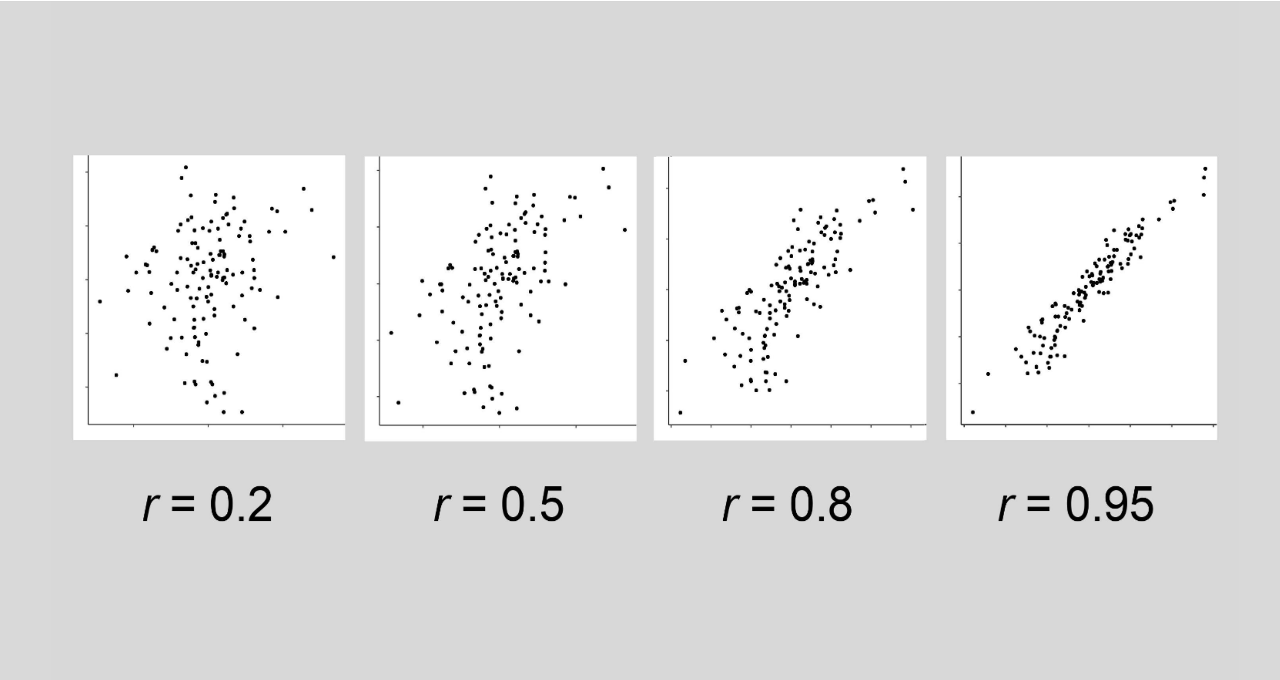

Figure 4.4. Participants viewed these plots for at least eight seconds before being allowed to continue to the practice trials.

4.5 General Methods

The experiments described in this chapter share multiple aspects of their procedures. Both experiments were built using PsychoPy [66] 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. [32] and discussed in Section 2.9 of Chapter 2 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 4.4); 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 4.5) 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. All 150 participants saw all 180 experimental items, corresponding to ~27,000 individual judgements per experiment, in a fully randomised order.

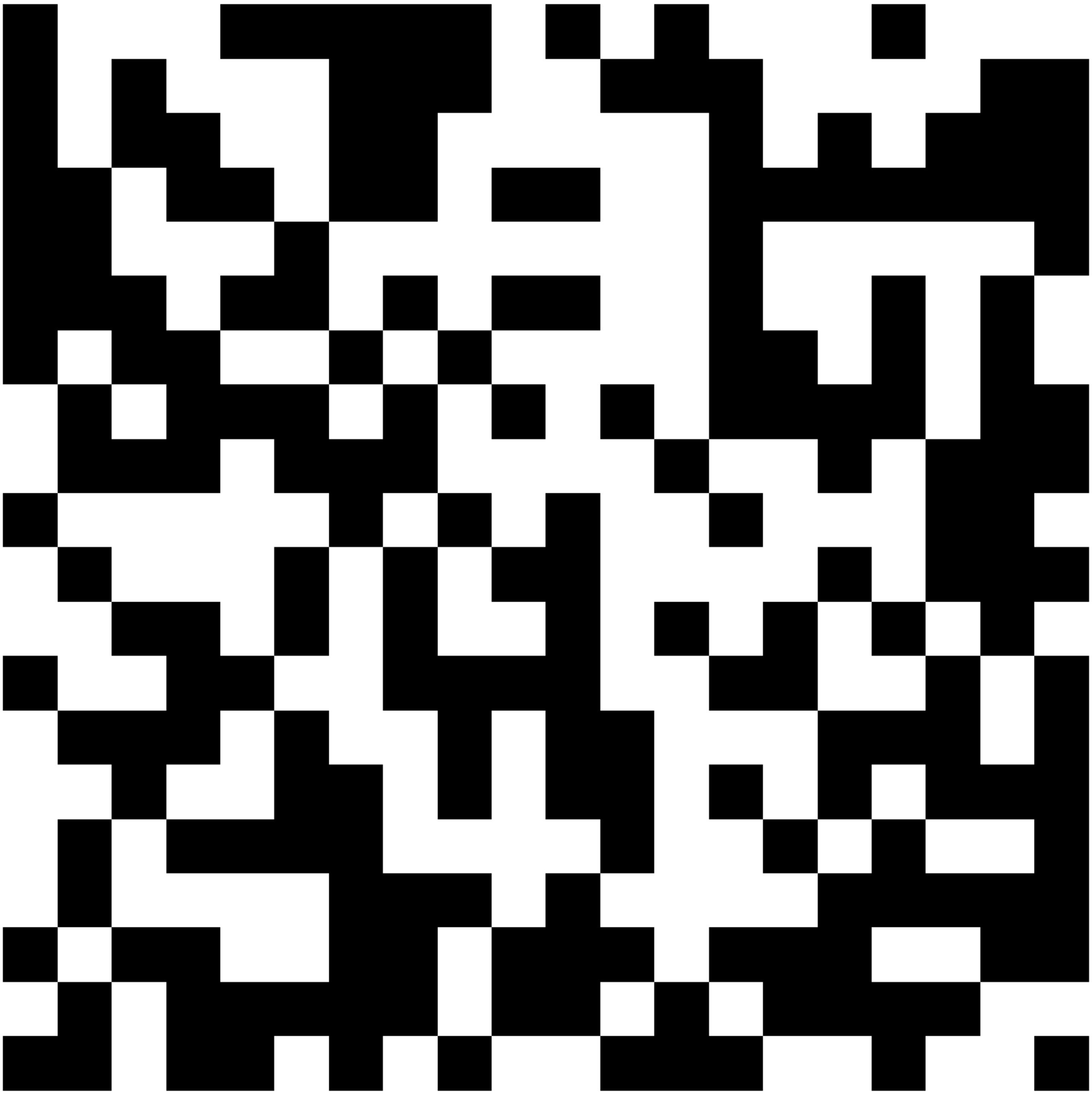

Figure 4.5. An example of a visual mask displayed for 2.5 seconds before each experimental trial.

Both experiments were conducted according to principles of open and reproducible research. All data and analysis code for the original paper are 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.

[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

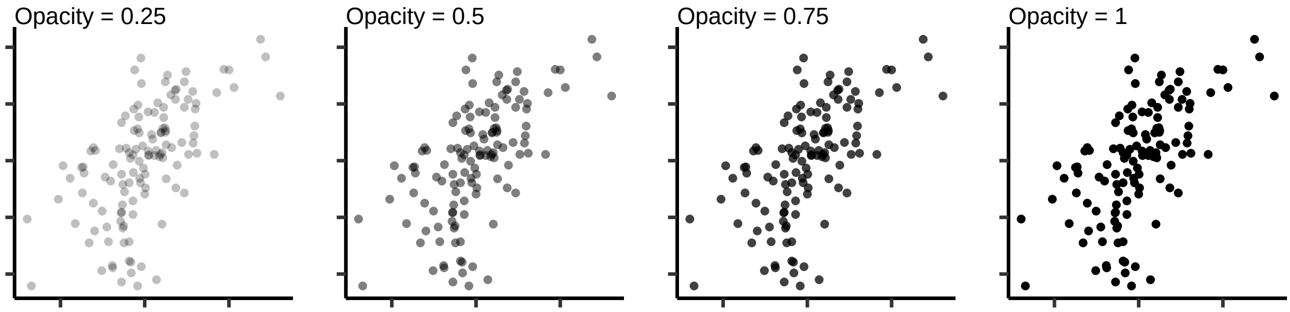

Figure 4.6. Examples of the stimuli used in experiment 1, demonstrated with an r value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.

4.6 Experiment 1: Uniform Opacity Adjustments

4.6.1 Introduction

Previous literature had described correlation perception as being resistant to changes in opacity [74, 75]. Findings from the pre-study described above provided evidence against this conclusion, so before proceeding with the fine-tuning of the immature point opacity decay function, I felt that gaining an understanding of how point opacity and correlation estimation interact more generally was important. Owing to the robust effects of altering stimulus opacity on perception described above [20, 98], it was hypothesised that:

- H1: A greater spread of estimates of correlation for plots with lower global opacity compared to higher opacity plots will be observed.

4.6.2 Method

Participants

150 participants were recruited using the Prolific platform [72]. 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 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 4.6, demonstrated with an r value of 0.6.

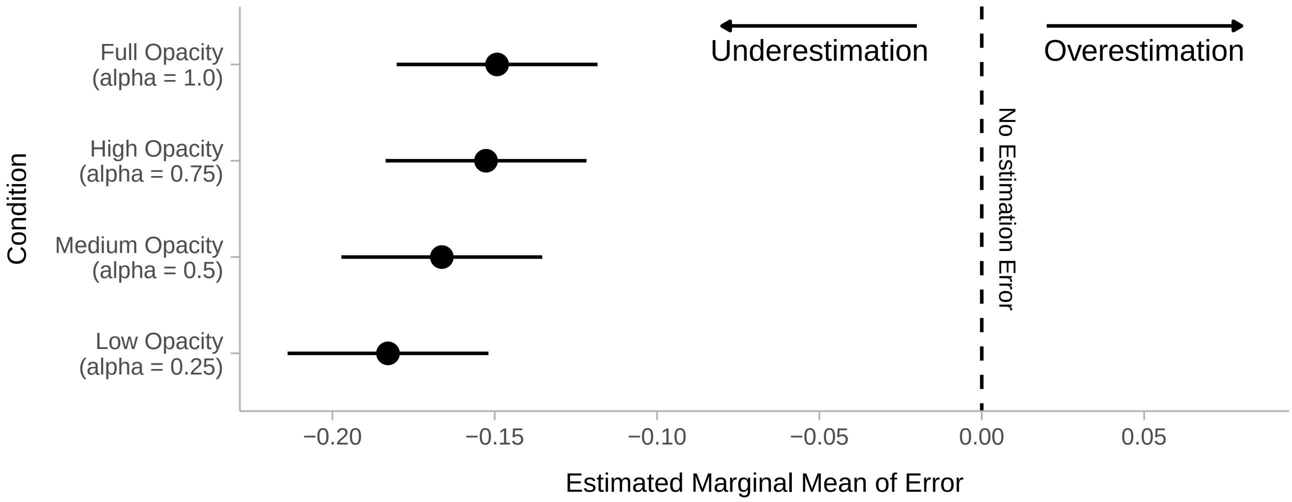

Figure 4.7. Estimated marginal means for the four conditions tested in experiment 1. 95% confidence intervals are shown. The vertical dashed line represents no estimation error. The overestimation zone is included to facilitate comparison to later work.

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

Contrast		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

4.6.3 Results

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 model including global opacity as a fixed effect explained significantly more variance than a null model ($\chi^2(3) = 223.13$, $p < .001$). Figure 4.7 shows the mean errors in correlation estimation for each opacity condition, along with 95% confidence intervals.

This effect was driven by significant differences between means of correlation estimation error between all conditions bar high and full opacity. Statistical tests for contrasts were performed using the `emmeans` package [47], 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 [41] 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

Table 4.4. Cohen's *d* effect sizes (left) and summary statistics (right) 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

Opacity	Mean	Standard Error
Full Opacity (alpha = 1.0)	0.15	0.016
High Opacity (alpha = 0.75)	0.15	0.016
Medium Opacity (alpha = 0.5)	0.17	0.016
Low Opacity (alpha = 0.25)	0.18	0.016

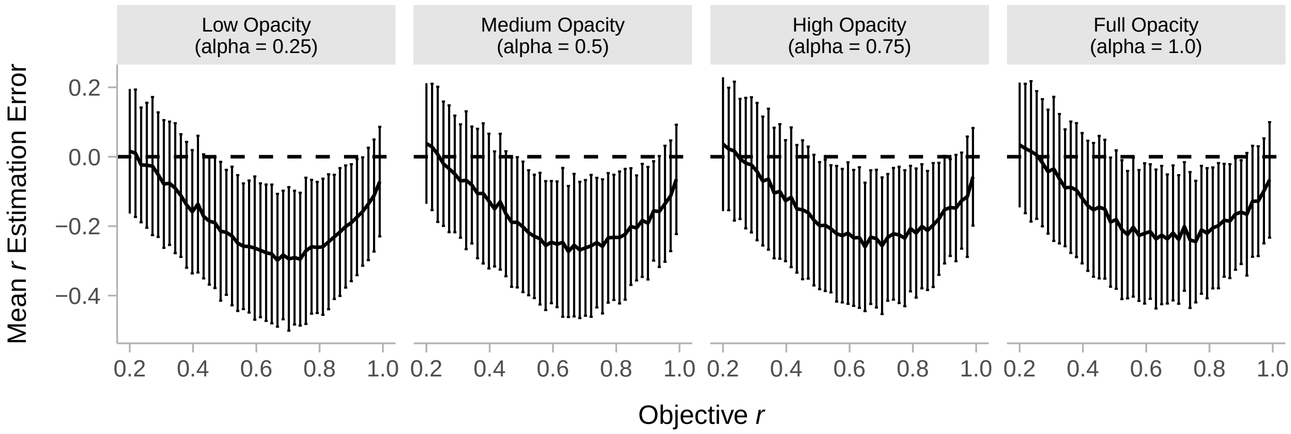

Figure 4.8. Participants' mean errors in correlation estimates grouped by factor and by *r* value. The dashed horizontal line represents perfect estimation. Participants were most accurate when presented with the plots featuring higher global point opacity. Error bars show standard deviations of estimates.

factor. These statistics can be seen in Table 4.4 along with means and standard deviations. 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 [74]. Figure 4.8 shows how participants' estimates of correlation change with the objective r value in the plot. Points represent mean errors in correlation estimation, and standard deviations of error are provided as error bars. The dashed horizontal line represents hypothetical perfect estimation. As reported in previous literature (see Sec in related work), participants underestimated r in nearly all cases (revise).

4.6.4 Discussion

The hypothesis, that there would be a greater spread of correlation estimates for plots with lower global opacity compared to those with higher global opacity, was not supported. As seen in Table 4.4 (right), standard errors for each opacity factor are identical to 3 decimal places. Participants' errors in correlation estimation were significantly greater when the opacity of all scatterplot points was lower compared when it was higher. This held true up until alpha was set to 0.75, implying a threshold around this value past which there is little variation in the perception of opacity, at least as far as it is associated with correlation estimation. This lack of significant difference in correlation estimation between the two highest global opacity conditions is congruent with the logarithmic nature of contrast/brightness perception [29, 94]; despite there being equal linear distance between the opacity values used, the perceptual distance between them was clearly non-linear.

As mentioned previously, Rensink [74, 75] present the only other account of experiments that directly test correlation perception as it pertains to the opacities of scatterplot points, and report no difference in either bias (error) or variability (spread) in correlation perception regarding point opacity manipulations. In comparison, the results observed here do report an effect. This effect may be explained by differences in experimental power, as it is a small effect, although I argue that methodological differences may have also played a small role. Given the small effect size (Cohen's $d = 0.16$), the small sample in Rensink (2014) [75] may have been insufficiently powered. With the large sample size in the present work (n = 150), evidence for an effect has been found. The experimental methodology utilised here is more representative of the use of scatterplots in the wild, and is therefore more suited to informing design as opposed to investigating the mathematical relationship between real and perceived correlation. While the effect is small, it demonstrates definitively that differences in the opacities of scatterplot points *can* affect estimates of correlation in positively correlated scatterplots. From the results it is unclear why lowering global opacity causes greater errors in correlation estimation while causing no difference in spread.

I suggest that correlation perception functions similarly to speed perception [20] with regards to changes in the contrast between foreground targets (in this case scatterplot points) and the background; the greater spatial uncertainty brought on by reduced point opacity, while not eliciting greater spread in correlation estimates, might be responsible for the effects observed via an increase in the perceived width of the probability distribution displayed by the scatterplot

From the results it is clear that a scatterplot optimised for correlation perception should have contrast between the foreground (points) and background in a range corresponding to alpha values of between 0.75 and 1. That there are significant differences in correlation estimation between data-identical scatterplots with different global point opacities however, suggests that this effect may be leveraged to further improve participants' performances on a correlation estimation task.

4.7 Experiment 2: Spatially-Dependent Opacity Adjustments

4.7.1 Introduction

Experiment 1 found that point opacity in positively correlated scatterplots has an effect on the perception of correlation such that those scatterplots with higher levels of global point opacity are rated as being more strongly correlated. Given this finding, the question arises of whether additional changes in correlation perception may be observed as a function of the spatial arrangement of point opacity. With this in mind, it was hypothesised that:

- H1: the non-linear decay parameter in which point opacity falls with residual distance will result in lower mean errors in correlation estimation compared to linear decay and full global opacity conditions.
- H2: the use of the inverted non-linear decay parameter, in which point opacity becomes greater with residual distance, will result in higher mean errors in correlation estimation than for all other conditions.

4.7.2 Methods

Participants

150 participants were recruited using the Prolific platform [72]. 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).

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 4.10, demonstrated with an *r* value of 0.6. Given the shape of the underestimation curve found in previous work (see related work chap 3 or something), intuition suggested employing a symmetrically opposing curve (see the non-linear decay curve in Figure 4.9) to relate point opacity to residuals.

Equation 4.1 was used to non-linearly map residuals to ggplot alpha values. 0.25 was chosen as the value of *b*, as it was felt at the time that this rendered plots that maintained point visibility while also allowing a large enough point opacity range that, if an effect was present, it was likely to be found.

$$point_{size/opacity} = 1 - b^{residual} \tag{4.1}$$

4.7.3 Results

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 model including opacity decay function as a fixed effect explained significantly more variance than a null model ($\chi^2(3) = 1{,}157.62$, $p < .001$). Figure 4.7 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 [47], and are shown in Table 4.5. To test whether the observed results could be explained by differences in graph literacy, a model including

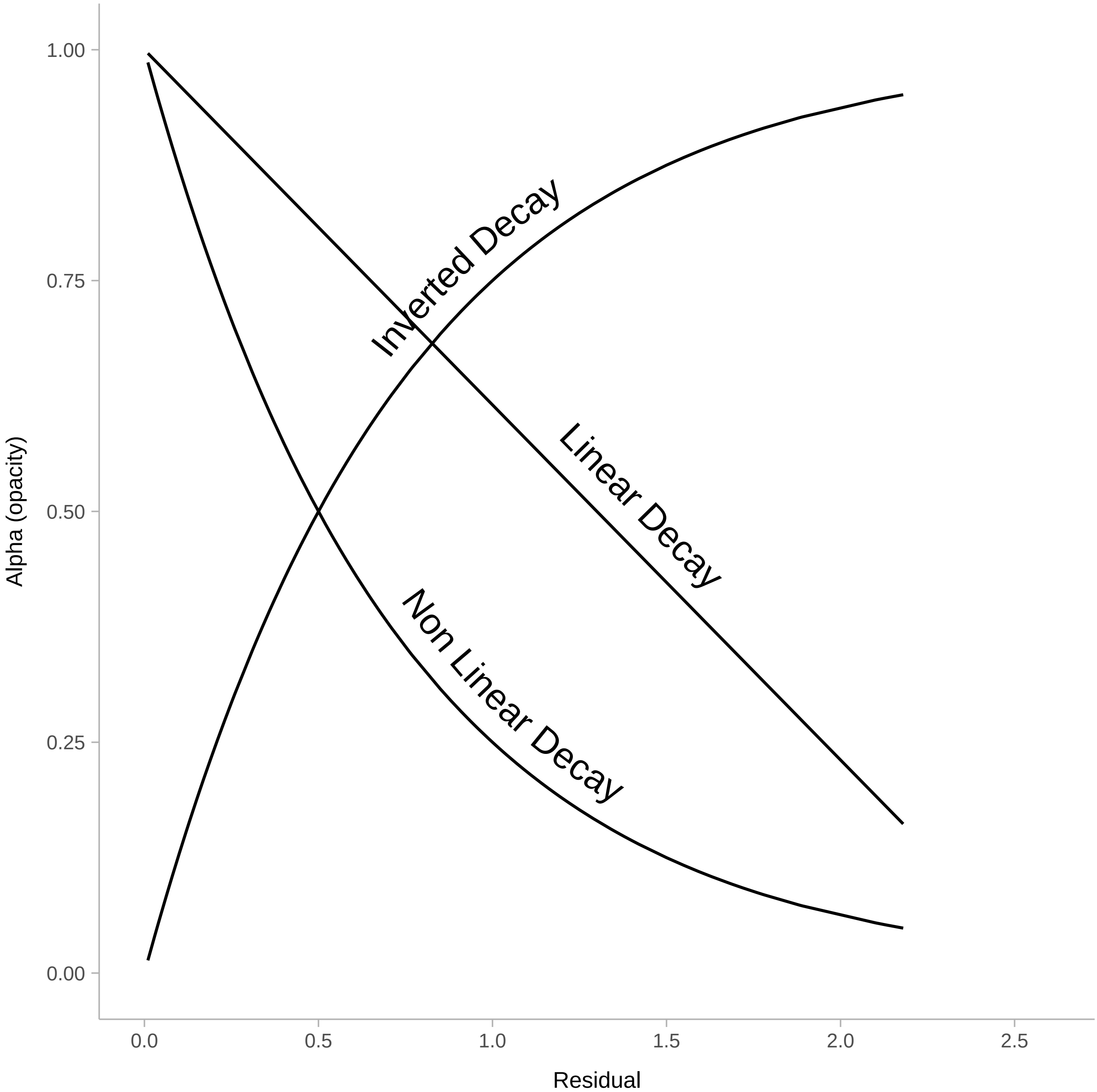

Figure 4.9. Using an r value of 0.2 to demonstrate the relationship between the size of a point's residual and the alpha value (opacity) rendered.

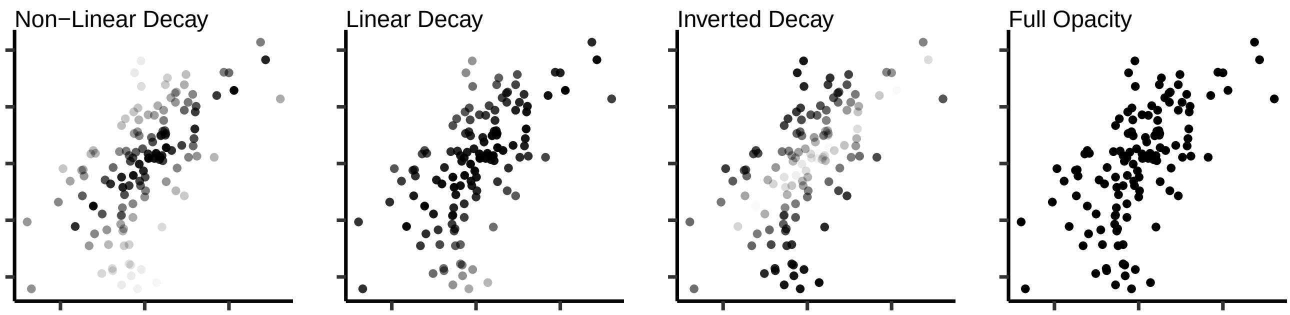

Figure 4.10. Examples of the stimuli used in experiment 2, demonstrated with an r value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.

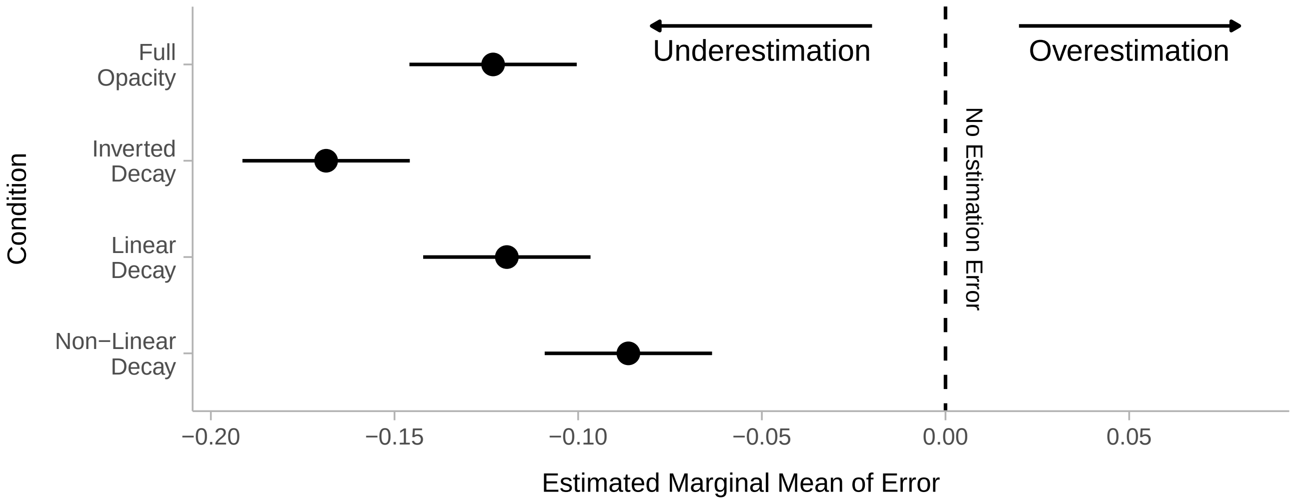

Figure 4.11. Estimated marginal means for the four conditions tested in experiment 2. 95% confidence intervals are shown. The vertical dashed line represents no estimation error.

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

Contrast		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

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. Figure 4.12 illustrates the effects of each manipulation on participants' correlation estimation performance separately for each value of *r* used. The dashed horizontal line represents perfect estimation, and standard deviations of estimation error are provided by way of error bars. Participants still underestimated correlation in all conditions, although the use of the non-linear decay function biased participants' estimates upwards to partially correct for the underestimation.

Table 4.6. Cohen's *d* effect sizes (left) and summary statistics (right) for 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

Opacity	Mean	Standard Error
Full Opacity	0.12	0.012
Inverted Decay	0.17	0.012
Linear Decay	0.12	0.012
Non-Linear Decay	0.09	0.012

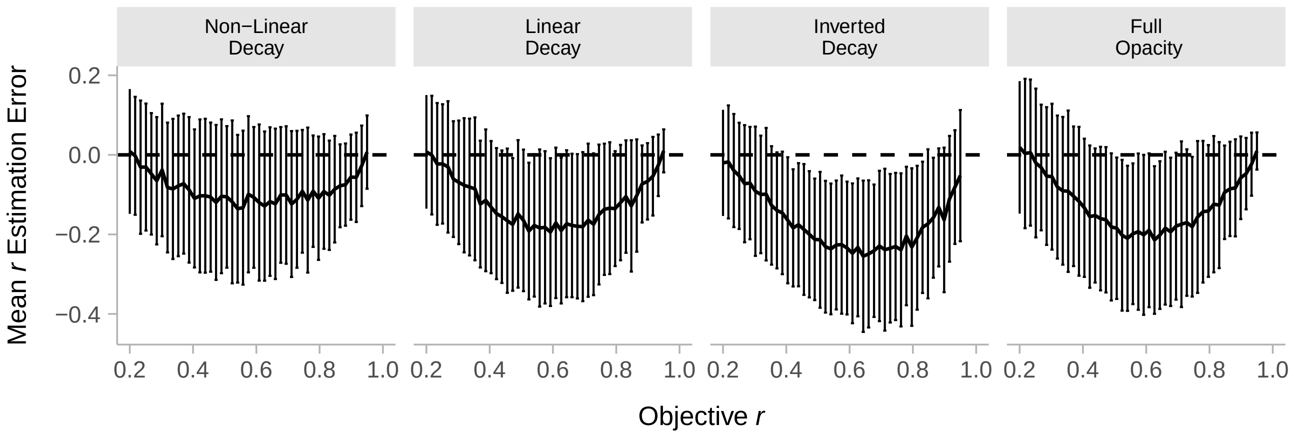

Figure 4.12. Participants' mean errors in correlation estimates grouped by factor and by *r* value. The dashed horizontal line represents perfect estimation. Participants were most accurate when presented with the plots featuring the non-linear opacity decay function. Error bars show standard deviations of estimates.

4.7.4 Discussion

Both hypotheses received support in experiment 2. Participants' errors in correlation estimation were lowest when the non-linear decay function was used, and were highest when scatterplots employed the non-linear inverted decay function. There was no significant difference in correlation estimation errors between the linear decay function and full contrast conditions. This result was surprising, however on closer inspection of the scatterplots in the linear decay function condition, it is clear that the logarithmic nature of contrast perception [29, 94] means there was little perceptual distance between points with high opacity values (*alpha* > 0.75). This resulted in no significant perceived differences between full opacity and linear decay function scatterplots. A similar threshold for high opacity values was found in experiment 1. Selecting only lower *r* values, those with naturally higher residuals (arbitrarily $r < 0.6$), still results in no difference between correlation estimation errors for linear decay parameters and full opacity conditions ($\chi^2(1) = 0.09$, $p = .769$). Effect sizes were small, with the largest being between the baseline full opacity and inverted non-linear decay conditions. This suggests that it is easier to induce further bias in correlation estimates through a reduction in the salience of a point cloud's centre than it is to correct for the underestimation bias.

Looking at the standard deviations of correlation estimates plotted separately by opacity decay function and *r* value in Figure 4.12, it can be observed that as in experiment 1 (see Figure 4.8), and aside from the inverted non-linear decay function condition, precision in *r* estimation increased with the objective *r* value. This finding corroborates previous work. In the inverted non-linear decay condition, as *r* approaches 1 (and point residuals accordingly approach 0), the opacity of points diminishes. Just as the standard deviation of correlation estimates was higher for the low global opacity condition in experiment 1, having lower opacity points at the high *r* end of the non-linear inverted condition in experiment 2 resulted in fairly constant standard deviations across *r* values, as the usual reduction towards $r = 1$ was not observed.

The non-linear inverted opacity decay condition produced significantly lower estimates of correlation than all other conditions. This adds perspective to suggestions [104] that, among other visual features, the area of a hypothetical prediction ellipse [24, 104], a region used to predict new observations assuming a bivariate normal distribution (see Figure 4.13) is a better predictor of correlation estimation

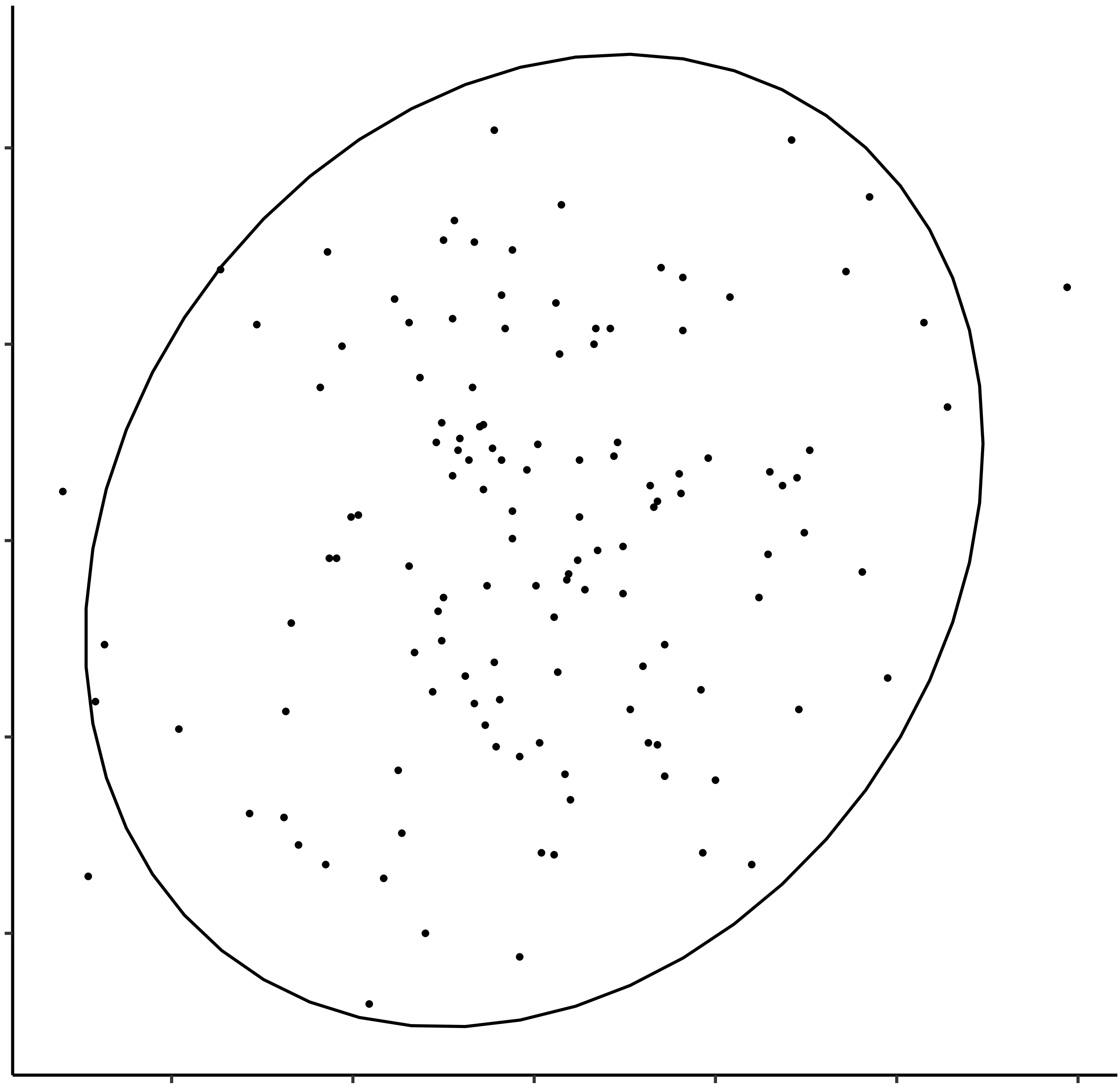

Figure 4.13. Plot showing a 95% prediction ellipse over a scatterplot with an r value of 0.6.

performance than the objective r value itself. In the inverted non-linear decay condition, the area of this ellipse remained the same, yet estimates of correlation differed significantly. These results suggest that the apparent density of the scatterplot point cloud also has effects on estimates of correlation. Prior work has found that more dense scatterplots are rated as having higher correlations [45, 75], although the effect found was weak. To explore this effect further, work investigating what people attend to when making correlation judgements must be completed. Eye-tracking offers an elegant solution to this problem, yet at the time of writing, had only been used for simpler scatterplot-related tasks, such as those asking participants to identify the number of, or distance between, points [60].

4.8 General Discussion

To summarise the results, evidence was found that changing the opacities of points in scatterplots can significantly alter participants' estimates of correlation, that lower global point opacity is associated with greater correlation underestimation error, that lowering point opacity as a function of residual error can partially correct for this underestimation, and that raising opacity as a function of residual error can increase the underestimation bias. These findings pave the way for using changes in point opacity to produce perceptually-optimised scatterplots that do not rely on data removal. As the focus of this work is on designing data visualisations that lay people are more easily able to interpret and understand, we used a large, representative sample, including people from a range of nationalities and educational backgrounds. This chapter demonstrates that a simple framework can be employed with these groups to gather high quality data, and, by design, produce conclusions that may generalise better to in-the-wild data visualisation usage.

In agreement with previous research [70, 74–77], participants were more accurate and precise in their estimates of correlation when the r value shown in the scatterplot was higher. Figure 4.8 and Figure 4.12 plot objective r value against participants' errors in correlation estimation separately for each level of the respective independent variable. These plots illustrate, as in much previous work, the lower levels of precision and accuracy in correlation estimation that are seen for r values further from 0 or 1.

The results here contribute to a body of evidence that suggests participants are attending to the width of the probability distribution displayed in a scatterplot [24, 55, 76, 104] when making judgements of correlation. Further evidence is provided for the systematic underestimation bias, and a potential correction strategy is offered. This work does not attempt to redesign the scatterplot as a medium, but to provide a set of recommendations for designers based on the evidence; when designing positively correlated scatterplots to support correlation perception:

- Lowering the total opacity in a scatterplot can cause people to underestimate correlation compared to when contrast is maximal between the points and the plot background (when point opacity is high).
- The use of a non-linear opacity decay function, in which point opacity falls as a function of residual size, can be used to counteract the underestimation seen in correlation estimation in positively correlated scatterplots.

Scatterplots are widely used, and are often designed with a number of communicative concepts in mind. When one of these concepts is illustrating the degree of positive association between two variables, the findings presented suggest that designers should utilise the techniques described to give viewers the best chance of interpreting the correlation displayed as accurately as possible.

4.8.1 Training

Both experiments tested lay participants with varying levels of graph literacy. Due to concerns about participants' familiarity with scatterplots, each saw four scatterplots depicting correlations of $r = 0.2$, 0.5, 0.8, and 0.95 (see Figure 4.4) to familiarise themselves with the concept. To test if the patterns of results seen in correlation estimation could be attributed to this training, models were built that included the half of the session (first or second) as a predictor for participants' judgements of correlation. Comparing these models to the original revealed a significant effect in experiment 1 ($\chi^2(1) =$

7.60, $p = 0.01$), but not experiment 2 ($\chi^2(1) = 2.20$, $p = 0.14$). In experiment 1, participants' errors in correlation estimation were higher in the second half of the experiment, suggesting that having more recently viewed the training material may have made participants more accurate. That this was not observed in experiment 2 implies that the point opacity manipulation had a greater effect on estimates than any training effects.

4.8.2 Limitations

The results in experiment 2 provide evidence that reducing the salience of points as they move further from the regression line can increase people's estimates of correlation, at least when plots like these are presented alongside conventional ones. Testing whether this phenomenon would exist with a plot in isolation would present a number of difficulties. As can be seen in Figure 4.8 and Figure 4.12, participants' estimates of correlation, especially between 0.2 and 0.7, suffer from high variance. High numbers of trials and participants ameliorate this to an extent, but this does prevent commentary on single plot judgements of correlation.

An important first step in the utilisation of point opacity changes to optimise perception of correlation in scatterplots has been made, however there remains much to do. Quantitative determination of whether 0.25 is indeed the most optimal value for b in equation 1, for example, is impossible from the present results. It may well be the case that changing the value of b as a function of the objective Pearson's r value could produce more accurate correlation estimation in participants; findings that participants were more accurate in correlation estimation when r was nearer 0 or 1 would suggest that the use of a decay parameter for these correlations is unnecessary.

The simplicity of the direct estimation task employed confers some limitations on the conclusions that can be drawn, although these limitations do not prevent the data gathered from being practically useful. The methods used to not allow for investigation of absolute correlation perception, as JND or bisection methodologies might. The finding that mean errors in judgements of correlation for full opacity conditions were different between experiments 1 (0.149) and 2 (0.123) makes this limitation evident. In the experimental paradigm, participants indirectly made comparative correlation judgements, which may have resulted in this discrepancy. Nevertheless, the results found are promising with regards to design research.

4.8.3 Future Work

Future work may wish to investigate negative values of r, given findings that people may overestimate the correlation of negatively correlated scatterplots in a similar way [83]. The techniques presented here could be adapted to bias perceptions of correlation down and correct for this overestimation. The non-linear opacity decay function demonstrated here may be able to accomplish this.

The opacity decay function conditions in experiment 2 used the vertical distance between a particular point and the regression line to set that point's opacity. Previous work [55] has suggested that the perpendicular distance between a point and the regression line may be a more accurate predictor of performance on correlation estimation tasks [24, 76, 104]. Future work exploring these manipulations further may wish to investigate whether there are differences between using perpendicular and vertical residual distance as bases for setting for opacity with regard to correlation estimation.

Finally, the most exciting avenue for related future work is the possibility of combining the techniques developed and tested here with other novel scatterplot visualisation techniques that have a basis in the literature. Chapter 5 investigates the use of point size in place of opacity with the techniques described, while Chapter 7 combines the two to explore how these different modalities work together with regards to the estimation of positive correlation.

4.9 Conclusion

In a pair of experiments, I varied the opacity of point in scatterplots both uniformly (experiment 1) and using functions relating point opacity to the size of a particular point's residual. In experiment 1, I showed that, in contradiction with previous work, changing the opacity of all points in a scatterplot can effect participants' performance on a correlation estimation task. In experiment 2, I showed that varying point opacity as a function of a point's residual distance is able to significantly change participants' correlation estimates and partially correct for a historic underestimation bias.

Chapter 5

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

5.1 Abstract

Chapter 4 provided strong evidence for the effects of systematically varying the opacities of scatterplot points on participants' estimates of correlation in positively correlated scatterplots. Utilising the same function and experimental paradigm, I show in a single experiment that systematically varying the sizes of scatterplot points is able to bias participants' estimates of correlation to a greater degree than manipulations that only adjust point opacity. In a condition where point size decreases non-linearly as a function of residual distance, correlation estimation is significantly biased upwards to correct for the historic underestimation bias to a greater degree. I discuss the implications of these findings for the mechanisms behind both opacity and size adjustments in scatterplots in relation to correlation estimation, and recommend techniques for those who design with the estimation of positive correlation in mind.

5.2 Introduction

While I was successful at changing participants' perceptions of correlation in positively correlated scatterplots in Chapter 4, the extent to which these perceptions were changed was minimal. Figure 4.12 illustrates how participants' mean errors in r estimation changed as a function of the subjective r value. Scatterplots employing non-linear opacity decay produced the most drastic changes in correlation estimation, however these changes were still small, with an effect size of Cohen's d = 0.19. Recent evidence suggests that with regards to altering percepts in scatterplots, changes in point size may be more effective than changes in opacity. In a fully-reproducible, large sample (N = 150) study, I show that systematically altering point size using the same function is not only able to more effectively correct for the correlation underestimation bias, but is also able to alter the shape of the correlation estimation curve.

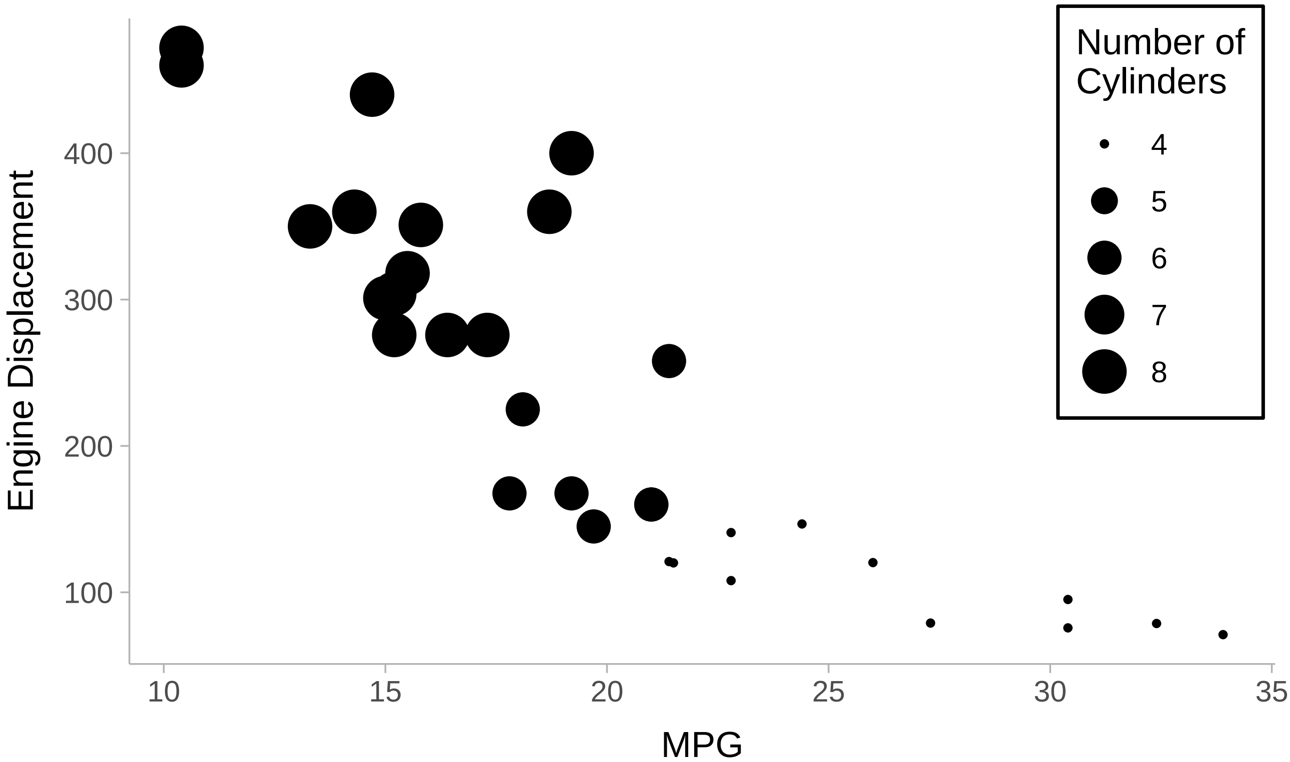

Figure 5.1. An example of a bubble chart. This plot compares car engine displacement (cubin inches) with fuel efficiency (miles per gallon). Additionally, the number of cylinders in the cars engine are encoded with point size. We can see from this plot that vehicles with higher engine displacement and lower fuel efficiency tend to have a higher number of engine cylinders.

5.3 Related Work

5.3.1 Point Size and the Perception of Correlation in Scatterplots

Opacity adjustments in scatterplots have been used extensively to solve issues of overplotting and clutter in scatterplots [12, 52]. Figure 4.2 in Chapter 4 demonstrates this usage. Practicality also dictates that scatterplots visualising sufficiently large datasets inherently require their points to be smaller to prevent obfuscation of the data they portray. Aside from being used to increase point visibility, point size changes in scatterplots have also been employed in the creation of bubble charts; here, the size of a scatterplot point is mapped to a third variable. Figure 5.1 demonstrates this class of scatterplot using the included `mtcars` dataset in `ggplot2`.

Despite the popularity of such charts, at the time of writing, very little investigation into how point size affects correlation perception in scatterplots had taken place. Those that were completed found bias and variability in correlation perception performance invariant to changes in point size [74, 75], however these studies were low-powered considering the small effect sizes found in the experiments described in Chapter 4. From the wider literature, there is evidence that larger points in scatterplots can bias judgements of mean point position to a greater degree than point opacity can [38], in what is termed the *weighted average illusion*. Outside of scatterplots and scatterplot-adjacent chart types, there is evidence that increased size can result in faster reaction times to peripherally presented stimuli [34], and there is evidence that larger stimuli are associated with lower levels of spatial certainty [2] but higher levels of salience [35]. The current experiment should therefore facilitate discrimination be-

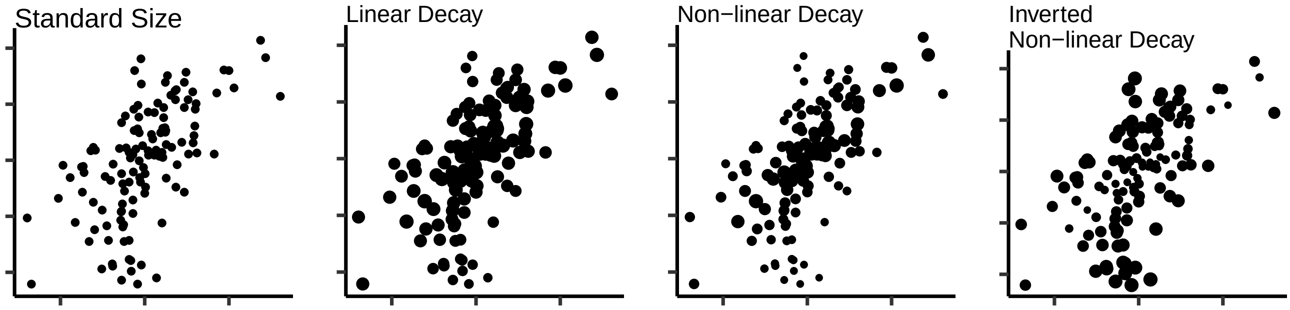

Figure 5.2. Examples of the stimuli used in experiment 3, demonstrated with an r value of 0.6.

tween these candidate drivers of the effects observed in a way that was not possible when manipulating opacity, as in that case effects of salience or spatial certainty would operate in the same direction.

5.4 Hypotheses

Based on the findings from Chapter 4, and on evidence from the wider literature described above, two hypotheses are made:

- H1: The non-linear decay parameter, in which scatterplots points become smaller as they move further from the regression line, will result in lower mean errors in correlation estimation than for linear decay and standard size conditions.
- H2: The use of an inverted non-linear decay parameter, in which scatterplot points become larger as they get further away from the regression line, will result in higher mean errors in correlation estimation than for all other conditions.

5.5 Methods

5.5.1 Stimuli

The creation of stimuli in this experiment follows the same general principles outlined in Section Chapter 3.2.4, Chapter 3. As in Chapter 4, equation 1 was used to map point residuals to size values in the two non-linear decay conditions:

$$point_{size/opacity} = 1 - b^{residual} \tag{5.1}$$

As in Chapter 4, the use of this equation produces a curve around the identity line symmetrically opposing the underestimation curve found in previous work. Additionally, a constant of 0.2 was added to each raw size value, and a scaling factor of 4 was utilised; these adjustments resulted in the smallest points in the present experiment having a width of 12 pixels on a 1920x1080 pixel monitor, which is consistent with the point size used in the experiments described in Chapter 4. Examples of the stimuli used in this experiment can be see in Figure 5.2.

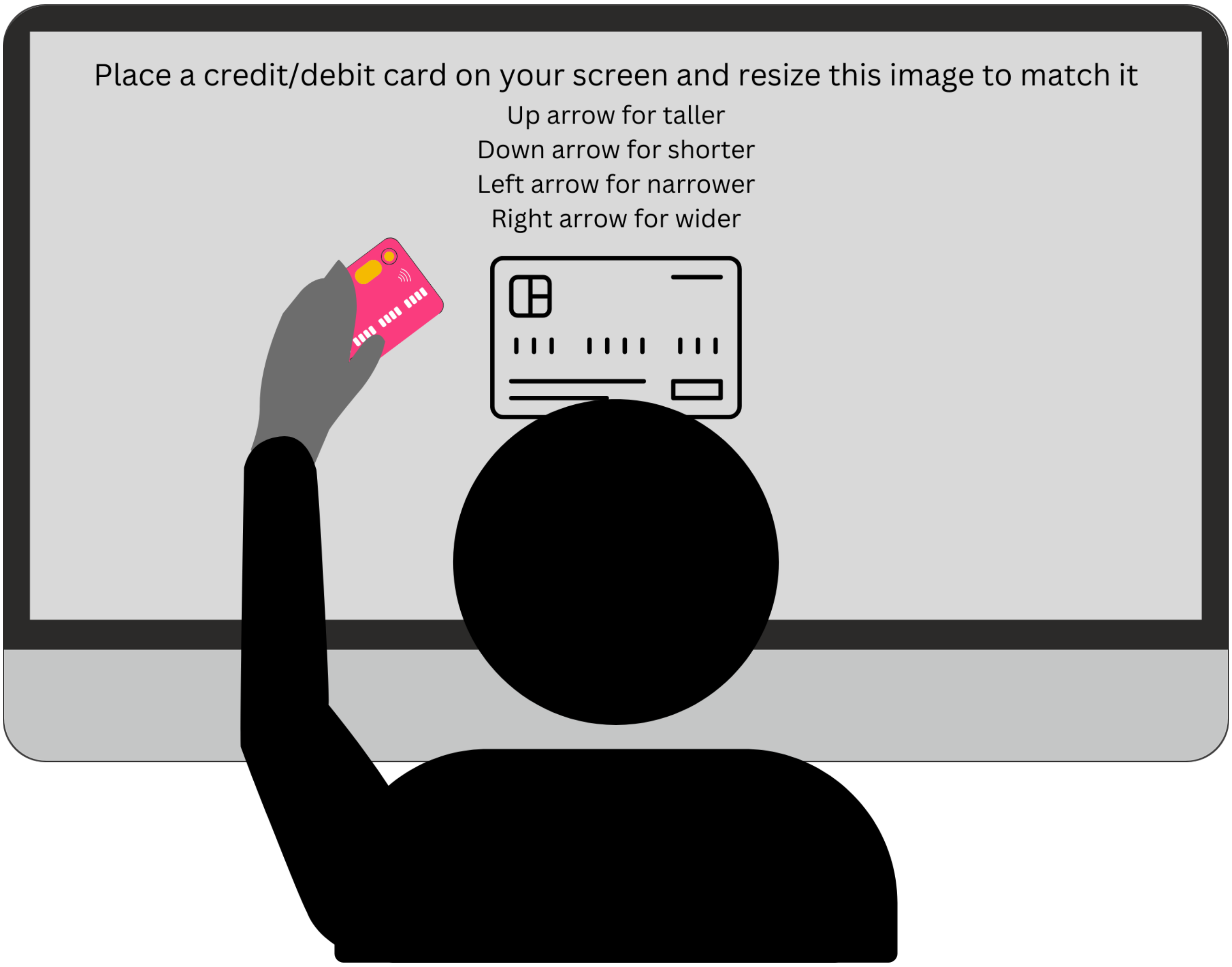

Figure 5.3. A mock-up of the screen scale [58] task used to infer the sizes of participants' monitors.

5.5.2 Dot Pitch in Crowdsourced Experiments

When the experiments described in Chapter 4 took place, no method of obtaining dot pitch was implemented. Dot pitch is defined as the distance between the dots (sub-pixels) [19] that make up each pixel. Calculating dot pitch is a requirement for the subsequent calculation of the physical on-screen sizes of the scatterplot points that participants saw. In the preamble to the current experiment, participants were asked to hold a standard size credit/debit/ID card up to the monitor, and then to resize a corresponding on-screen image until it matched the size of their physical card [58]. These cards have a universal standard size (ISO/IEC 7810 ID-1), which when combined with the monitor resolution information recorded by Psychopy, and assuming a widescreen 16:9 aspect ratio, allows for the inference of dot pitch and therefore the physical size of the points in the experience. Mean dot pitch was 0.33mm ($SD = 0.06$), corresponding to a physical size on the screen of 4.32mm for the smallest points displayed. Section 5.6 includes analysis that takes into account the physical on-screen sizes of scatterplot points.

5.5.3 Point Visibility Testing

It is key that the manipulations used do not remove (or appear to remove) data from scatterplots. Therefore, point visibility testing is included in this experiment prior to the experimental items. Participants were shown 6 scatterplots and were asked to enter in a text box how many points were being displayed. These points were the same size as the smallest points displayed in the experimental items. 5% of participants were correct on 5 out of 6 point visibility tests, while 95% were correct on 6 out of 6. It should be noted that those participants scoring 5/6 did not answer incorrectly, rather they did not answer at all for a particular question, which is suggestive of a mis-click or an initial misunderstanding of the task. Regardless, the results of this test indicate a sufficient level of point visibility.

5.5.4 Design

Again, a fully repeated-measures, within-participants design was employed. Each participant saw and responding to each of the 180 scatterplots in a fully randomised order. There were four scatterplots for each of the 45 r values corresponding to the four levels of the size decay condition, examples of which are shown in Figure 5.2. The experiment itself is hosted on Pavlovia [1].

5.5.5 Procedure

Each participant viewed the PIS and provided consent through key presses in response to consent statements. Participants were asked to provide their age in a free text box, followed by their gender identity. Participants then completed the 5-item Subjective Graph Literacy test [32], followed by the screen scale and point visibility tasks described above. Participants were shown example of scatterplots depicting r values of 0.2, 0.5, 0.8, and 0.95. Section 5.6 contains discussion of the potential effects of this training. Following two practice trials, participants worked through the series of 180 experimental and six attention check trials in a randomised order. Visual masks preceded each plot.

5.5.6 Participants

Normal to corrected-to-normal vision and English fluency were required. Participants who had completed any of the experiments described in Chapter 4 were prevented from participating. Data were collected from 164 participants. 14 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 (48% male, 50 % female, and 2% non-binary). Participants' mean age was 29.56 (SD = 8.54). Mean graph literacy score was 21.77 (SD = 4.29). The average time taken to complete the experiment was 39 minutes (SD = 14 minutes).

5.6 Results

To investigate the effects of point size adjustments on participants' estimates of correlation, a linear mixed effects model was built whereby the point size 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 participants and items. A likelihood ratio test revealed that the model including size decay function as

[1] https://gitlab.pavlovia.org/Strain/exp_size_only

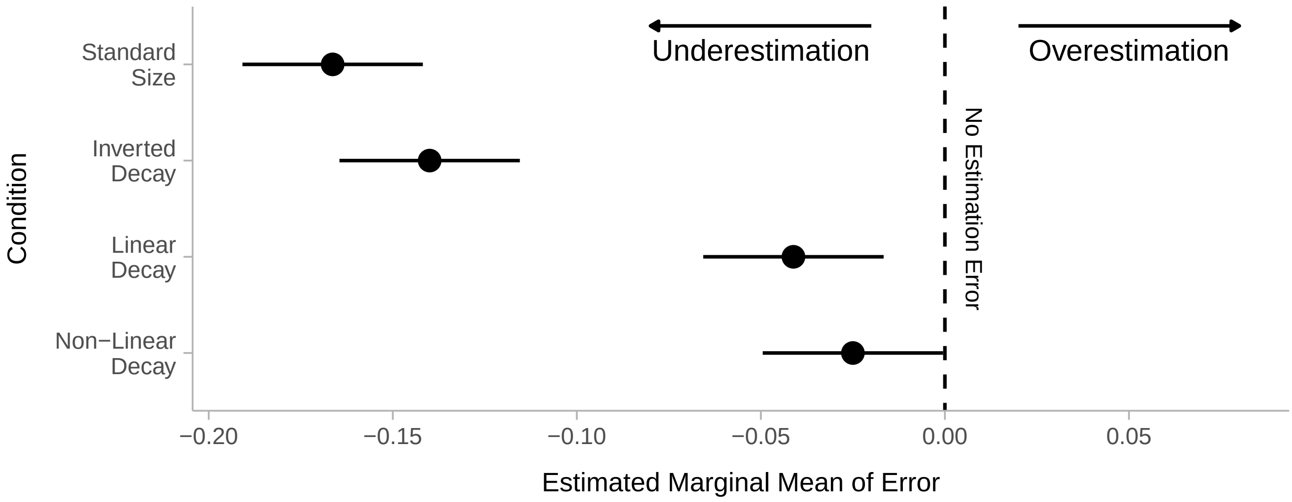

Figure 5.4. Estimated marginal means for the four conditions tested in experiment 3. 95% confidence intervals are shown. The vertical dashed line represents no estimation error. The overestimation zone is included to facilitate comparison to later work.

Table 5.1. Contrasts between different levels of the size decay factor in experiment 3.

Contrast		Statistics	
		Z ratio	*p*
Standard Size	Inverted Decay	9.3	<0.001
Standard Size	Linear Decay	44.4	<0.001
Standard Size	Non-Linear Decay	50.1	<0.001
Inverted Decay	Linear Decay	35.1	<0.001
Inverted Decay	Non-Linear Decay	40.8	<0.001
Linear Decay	Non-Linear Decay	5.7	<0.001

a fixed effect explained significantly more variance than the null model ($\chi^2(3) = 3{,}508.84$, $p < .001$). Figure 5.4 shows the mean errors in correlation estimation for each size decay condition, along with 95% confidence intervals.

This effect was driven by significant differences between means of correlation estimation error between all conditions. Statistical tests for contrasts were performed using the `emmeans` package [47], and are shown in Table 5.1. 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 ($\chi^2(1) = 0.16$, $p = .690$), indicating that the differences observed in participants' correlation estimation performance were not as a result of differences in levels of graph literacy.

While participants performed well on the point visibility task, another facet of using a larger or smaller monitor with a lower or higher resolution could have affected estimates of correlation. Comparing a model including the dot pitch of participants' monitors to the experimental model revealed a significant main effect ($\chi^2(1) = 4.65$, $p = .031$). There was no interaction between size decay condition and dot pitch; a 0.1mm decrease in dot pitch resulted in correlation estimates decreasing by .03. Given the low range of dot pitches gathered from participants 0.13mm to 0.63mm, the effect is not substantial enough to warrant further discussion.

Table 5.2. Cohen's d effect sizes (left) and summary statistics (right) for levels of the size decay factor in experiment 3. Each effect size is compared to the reference level, termed, "Standard Size".

Effect	Cohen's d
Standard Size	
Inverted Decay	-0.11
Linear Decay	-0.54
Non-Linear Decay	-0.61

Size	Mean	Standard Error
Standard Size	0.17	0.013
Inverted Decay	0.14	0.013
Linear Decay	0.04	0.013
Non-Linear Decay	0.02	0.013

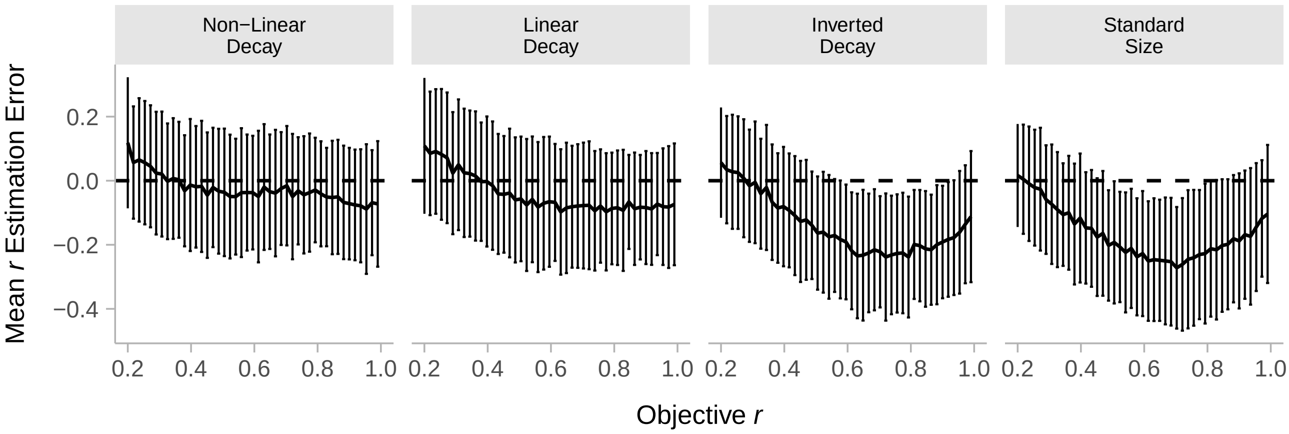

Figure 5.5. Participants' mean errors in correlation estimates grouped by factor and by r value. The dashed horizontal line represents perfect estimation. Participants were most accurate when presented with the plots featuring the non-linear size decay function. Error bars show standard deviations of estimates.

As in Chapter 4, an approximation of Cohen's d was calculated between the reference level (here labelled "Standard Size") and all other levels of the size decay factor. These statistics can be seen in Table 5.2 along with descriptive statistics for all levels. The largest observed effect size is between the standard size and non-linear decay conditions (d = 0.64), and is medium. This is a marked improvement on the non-linear opacity decay condition from experiment 2, Chapter 4, where the effect size observed was d = 0.19. Figure 5.5 plots the effects of each size decay manipulation separately for each r value used. Points represent means of estimation error for each r value, with standard deviations of error shown as error bars. The dashed horizontal line represents hypothetical perfect estimation. Expectedly, while participants still underestimated r in all conditions, the non-linear size decay condition provided the most promising correction to date; the average estimation error is approaching a flat line (see the left-most plot in Figure 5.5).

5.7 Discussion

Participants' errors in correlation estimation were significantly lower in the non-linear size decay condition (see Figure 5.5) compared to all other conditions. This finding provides support for the first hypothesis. Conversely, no support was found for the second hypothesis, that estimates would be least accurate in the inverted non-linear size decay condition. Errors in this condition were significantly higher than for the other two decay conditions, but were significantly lower than the errors that were observed for the standard size condition.

5.7.1 Increased Correlation Estimation Accuracy

The mean error in correlation estimation for the non-linear size decay condition in this experiment was .025, while for the equivalent opacity decay condition described in experiment 2 (see Chapter 4) was .086. This finding, along with the significantly higher Cohen's *d* effect size reported in this experiment compared to experiment 2 for the non-linear decay conditions (0.86 vs 0.19), provides evidence that point size is a stronger encoding channel for the manipulation of perceived contrast than point opacity. If these effects are being driven by the lower salience of more external points, the fact that a larger effect of point size has been reported is congruent with research showing clear influences of stimulus size on object salience and perceptual weighting [34, 35, 38]. The present results therefore provide support for point salience/perceptual weighting being a key driver of the effects observed; lower point salience brought on by reduced point size in the exterior of the scatterplot reduces the perceived width of the distribution of data points around the regression line, biasing estimates of correlation upwards and leading to a higher degree of accuracy. Other candidate mechanisms do exist; similar results would be expected if a feature-based attentional bias was responsible [38, 90]; the current methodology does not allow for distinguishing between these explanations, and it may be that both are partially responsible.

The lack of support for the second hypothesis is surprising, and suggests that point salience and perceived distributional width do not form the whole story. There is evidence that larger stimuli exhibit greater levels of spatial uncertainty [2], and it is possible that this uncertainty results in a perceptual under-weighting of contribution of these points during correlation estimation. This is consistent with previous work [96, 97] suggesting that the brain may make robust statistical use of visuo-spatial information. These mechanisms act to downweight the influence of less reliable information (in this case the higher spatial uncertainty of larger exterior points) on subsequent perceptual estimates. In the present experiment, this resulted in participants making more accurate estimates of correlation in the inverted decay condition compared to the standard size condition. It was suggested in Chapter 4 that inverted opacity manipulations may be employed to correct for the *overestimation* of correlation observed with negatively correlation scatterplots [83]; findings here indicate that using an inverted size decay function in this way may not be appropriate.

5.7.2 Constant Correlation Estimation Precision

Unlike the experiments described in Chapter 4, in which standard deviations of correlation estimation errors generally became smaller as the actual r value increased, distributions of standard deviations here remained mostly constant. This was unexpected, as previous work [74–77] routinely finds precision in r estimation to increase with the objective r value. This finding may be related to the nature of the stimuli in the current study. At high values of r there is a large amount of overlap between points in the non-linear, non-linear inverted, and linear size decay conditions. This overlap may blur the percept and account for the absence of this effect, however cannot account for these findings with regards to the standard size condition. Aside from the inverted non-linear decay condition in Chapter 4, the finding that precision increased with r was robust. Its absence here is curious given that the standard size decay condition here is identical to the full opacity condition in that chapter. Relying on relative judgements means the interplay between scatterplots with different visual features must be accounted for within a particular experiment. The stimuli as r approaches 1 in the current study exhibit greater levels of visual variance than the stimuli in experiments 1 and 2 (see Chapter 4), which may explain the lack of increased precision here. Further testing is required for a more concrete explanation.

Ultimately, my aim is to provide tools for the design of visualisations more suited for the tasks they are intended to support. When that task is the perception of positive correlation, the use of the non-linear size decay condition described here is recommended. For other scatterplot tasks, such as cluster separation or numerosity perception, or other chart types, the use of the size manipulation may in fact be a hindrance.

5.7.3 Training

Before beginning the experimental trials, participants viewed plots depicting r = 0.2, 0.5, 0.8, and 0.95 for a minimum of 8 seconds. Comparing a model including session half as a fixed effect with the origin experimental model revealed no significant effect ($\chi^2(1) = 1.06$, $p = 0.30$), suggesting that having more recently viewed the example plots did not have an effect on participants' performance.

5.7.4 Limitations

Despite confirming a method of obtaining dot pitch, no method of obtaining head-to-monitor distances is available. This, along with the comparative correlation judgements collected, prevents concrete psychophysical conclusions from being made. Instead, the experimental paradigm allows for findings that are rigorous to different viewing contexts and are of particular importance for the HCI and visualisation design audiences. It may be that a high level perceptual phenomenon is responsible for the effects have seen here; investigating this is beyond the scope of the current study and does not negate the findings. There is also the potential for misinterpretation of the scatterplots presented in the current study, especially given their similarity in form, but not purpose to bubble charts.

5.7.5 Future Work

Evidence has been found for robust effects of varying point opacity and point size on correlation estimation in positively correlated scatterplots. There are different effects of changing these visual features, and these effects are also partially dependent on the objective r value in the scatterplot. Future work should investigate the combination of these visual manipulations as they pertain to correlation estimation.

Additionally, future qualitative work is required to investigate whether the modified scatterplots presented result in misinterpretation or a decrease in the levels of trust people place in data visualisations.

5.8 Conclusion

I conducted a single experiment, in which points on scatterplots had their sizes systematically changed according to their distance from the regression line. The equation relating point size to residual distance was identical to that used in Chapter 4, with the addition of a scaling factor and constant to provide parity between the smallest points in the previous experiment and those in the current. Additionally, I tested the visibility of the smallest points used to ensure that no data were functionally removed, as well as collecting the sizes and resolutions of participants' monitors. I found evidence that reducing the size of scatterplot points as a function of residual error using a non-linear equation not only producing significantly more accurate estimates of positive correlation, but also subtly changed the fundamental shape of the estimation curve in a way not previously reported. Unlike the inverted opacity decay function used in Chapter 4, the inverted size decay function here did not produce the least accurate estimates of correlation; these were found when participants viewed the "standard size" plot devoid of a size manipulation. Taken together, these results suggest a greater potential for manipulating correlation estimates when using point size compared to point opacity.

Chapter 6

Interactions of Opacity and Size Adjustments

6.1 Abstract

6.2 Introduction

6.3 Related Work

6.4 Hypotheses

A single experiment is present in this chapter based on the effects of adjusting point opacity and size on correlation estimation established in Chapter 4 and Chapter 5. Here, previously independently tested point opacity and size manipulations are tested in both typical orientation (point opacity/size is reduced with residual magnitude) and inverted orientation (point opacity/size is increased with residual magnitude). Throughout this chapter, referral is made to *congruent* and *incongruent* conditions with respect to the combination of point opacity and size decay functions. *Congruent* conditions are those in which point opacity and size decay act in the same direction (typical or inverted), while for *incongruent* conditions, point opacity and size decay act against each other. Due to previous findings that non-linearly reducing point opacity or size as a function of residual magnitude can bias correlation estimates upwards, it is hypothesised that:

- H1: an increased reduction in correlation estimation error will be observed when congruent typical orientation decay functions are used. H2: the use of a congruent inverted function, such that point opacity and size are both increased with residual magnitude, will produce the least accurate estimates of correlation.

Owing to the finding from Chapter **??** that point size is a stronger channel for biasing correlation estimation than point opacity, it is also hypothesised that:

- H3: there will be a significant difference in correlation estimates between the two incongruent orientation conditions.

Table 6.1. Contrasts between different levels of the opacity and size decay factors in experiment 4.

Contrast		Statistics	
		Z ratio	*p*
TO Size x IO Opacity	IO Size x IO Opacity	-10.945	<0.001
TO Size x IO Opacity	TO Size x TO Opacity	72.294	<0.001
TO Size x IO Opacity	IO Size x TO Opacity	-2.256	0.108
IO Size x IO Opacity	TO Size x TO Opacity	46.125	<0.001
IO Size x IO Opacity	IO Size x TO Opacity	17.838	<0.001
TO Size x TO Opacity	IO Size x TO Opacity	-37.438	<0.001

6.5 Methods

6.5.1 Stimuli

6.5.2 Design

6.5.3 Procedure

6.5.4 Participants

Normal to corrected-to-normal vision and English fluency were required. Participants who had completed any of the experiments described in Chapter 4 or Chapter 5 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 from the remaining 150 participants were included in the full analysis (51% male, 49 % female, and 1% non-binary). Participants' mean age was 30.65 (*SD* = 8.64). Mean graph literacy score was 22.49 (*SD* = 3.55). The average time taken to complete the experiment was 37 minutes (SD = 12.3 minutes).

6.6 Results

To investigate the effects of combining point opacity and size decay functions on participants' estimates of correlation, a linear mixed effects model was built whereby the particular combination of point opacity and size decay function employed is a predictor for the difference between objective *r* values for each plot and participants' estimates of *r*. Deviation coding was used for each of the experimental factors, which allows comparison between means of *r* estimation error and the grand mean. This model has random intercepts for items and participants, and random slopes for participants with regards to the size decay factor. A likelihood ratio test revealed that the model including the opacity and size decay conditions as predictors explained significantly more variance than the null ($\chi^2(3)$ = 5,286.81, p < .001). There were significant fixed effects of opacity and size decay function, as well as a significant interaction between the two. Figure 6.1 shows the mean errors in correlation estimation for each combination of conditions, along with 95% confidence intervals.

The effects found were driven by significant difference between means of correlation estimation error

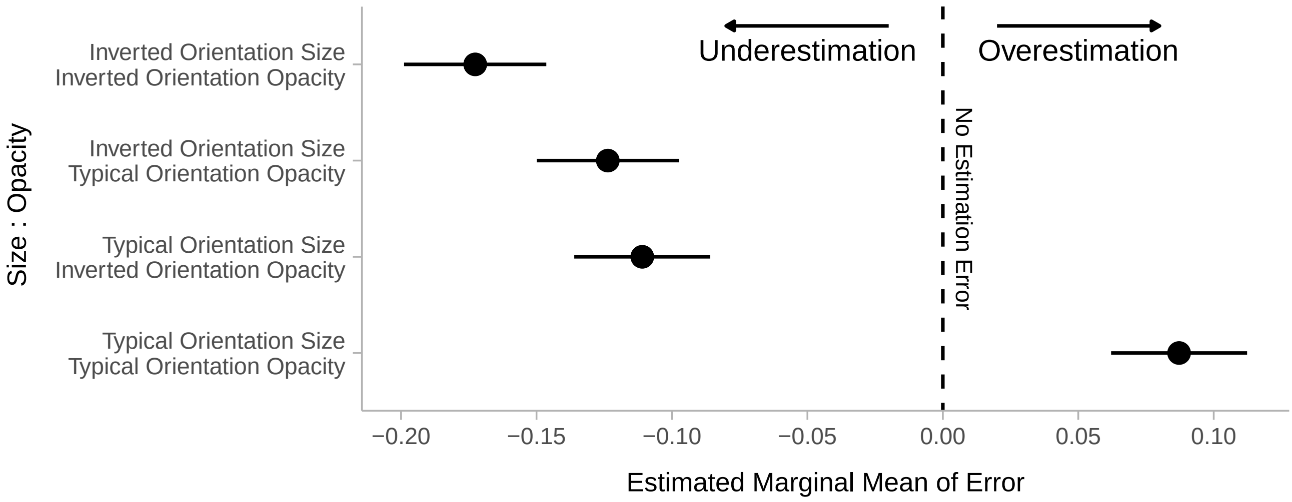

Figure 6.1. Estimated marginal means for the four conditions tested in experiment 4. 95% confidence intervals are shown. The vertical dashed line represents no estimation error.

Table 6.2. Significances of fixed effects and the interaction between them. Semi-partial R^2 for each fixed effect and the interaction term is also displayed in lieu of effect sizes.

	Estimate	Standard Error	df	t-value	*p*	R^2
(Intercept)	0.08	0.013	103.32	6.27	<0.001	
Size Decay	-0.14	0.005	148.39	-25.77	<0.001	0.104
Opacity Decay	0.12	0.002	26327.21	63.71	<0.001	0.087
Size Decay x Opacity Decay	0.15	0.004	26327.13	38.47	<0.001	0.034

between all conditions besides that which compares the two incongruent decay conditions. Statistical testing for contrasts were performed using the `emmeans` package [47], and are provided in Table 6.1. Experiments 1, 2, and 3 all featured a comparative baseline condition. In the former two experiments, this was the full contrast condition (see Chapter 4), while in the latter, this was the standard size condition (see Chapter 5). In the current experiment, no baseline condition was used. Owing both to this and the use of a linear mixed effects model with an interaction term, the use of Cohen's *d* as a measure of effect size would be inappropriate. In its place, the amounts of variance in participants' errors in correlation explanation explained by each fixed effect term and the interaction term is represented as semi-partial R^2 [59]. These statistics were calculated using the `r2glmm` package (version 0.1.2) [39], and are presented along with model statistics in Table 6.2.

Models including participants' graph literacy, their performance on the point visibility task, the dot pitch of participants' monitors, and which half a particular correlation judgement took place were built and compared with the experimental model. While no significant effects of graph literacy ($\chi^2(1)$ = 3.50, p = .061), performance on the point visibility task ($\chi^2(1)$ = 1.29, p = .257), or dot pitch ($\chi^2(1)$ = 1.52, p = .218) were found, there was a significant effect of training ($\chi^2(1)$ = 23.78, p < .001), with participants rating correlation .01 lower during the second half. This drop suggests that having more recently viewed the training plots may have increased participants estimates of correlation. To further analyse this variability, a model was built including trial number, allowing for the analysis of error. A significant effect of trial number is also found ($\chi^2(1)$ = 29.31, p < .001) on participants' correlation estimation errors.

Figure 6.2 shows participants' unsigned mean errors in correlation estimation against trial number.

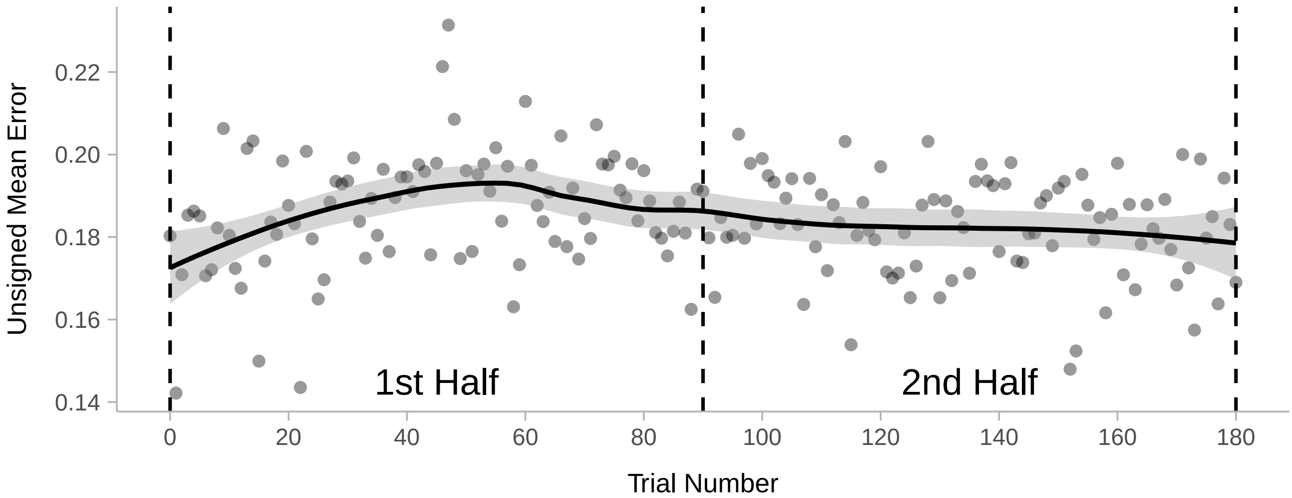

Figure 6.2. Comparing mean errors in correlation estimation by trial number. Points represent unsigned mean errors for each trial number. The plotted line is the locally estimated smoothed curve, with the ribbon representing standard errors.

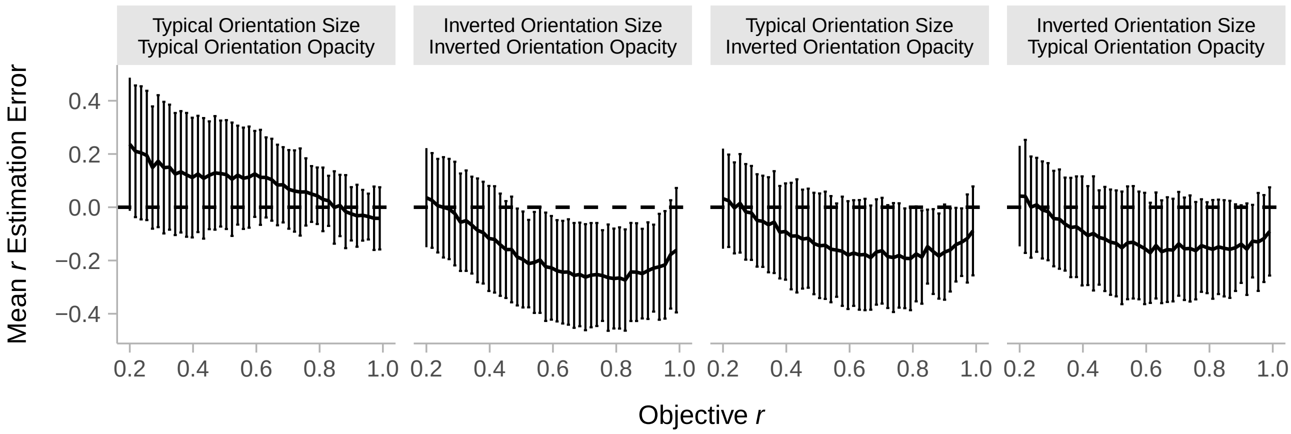

Figure 6.3. Participants' mean errors in correlation estimates grouped by factor and by r value. The dashed horizontal line represents perfect estimation. Participants were most accurate when presented with the plots featuring the non-linear size decay function. Error bars show standard deviations of estimates.

Variability in error, as represented by the ribbon, stabilised quickly and remained stable for most the experiment, only widening again around trial number 170. The simplest explanation for this is that participants, knowing they were coming to the end of the experiment, became less vigilant and rushed their judgements more. Regardless of statistical significance, this effect is not large enough to warrant further investigation, at least as it pertains to correlation estimation in scatterplots.

6.7 Discussion

Hypothesis 1 received full support in this experiment. The combination of typical orientation opacity and size decay functions produced the most accurate estimates of correlation, although this also resulted in a marked over-correction and consequent overestimation for many values of r (see Figure 6.3). The second hypothesis also received support; the combination of inverted opacity and size decay functions produced the least accurate estimates of correlation. No support was found for the third hypothesis, that there would be a significant difference in correlation estimates between inverted orientation opacity/typical orientation size plots and typical orientation opacity/inverted orientation size plots. There was, however, a significant interaction term present, providing evidence that the combination of opacity and size decay functions is not additive in natures.

6.7.1 Combining Manipulations

6.7.2 Estimation Precision

6.7.3 Relative Contributions of Opacity and Srize Decay

6.7.4 Mechanisms

6.7.5 Limitations

6.7.6 Future Work

6.8 Conclusion

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. 29).

[2] David Alais and David Burr. "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration". In: *Current biology: CB* 14.3 (Feb. 2004), pp. 257–262. ISSN: 0960-9822. DOI: `10.1016/j.cub.2004.01.029` (cited on pp. 53, 59).

[3] JJ Allaire and Christophe Dervieux. *Quarto: R Interface to 'quarto' Markdown Publishing System*. Manual. 2024 (cited on p. 27).

[4] 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. 29).

[5] 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. 20).

[6] 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. 22).

[7] 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. 20).

[8] 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. 14).

[9] 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. 26).

[10] 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. 26).

[11] 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. 26).

[12] Enrico Bertini and Giuseppe Santucci. “Quality Metrics for 2D Scatterplot Graphics: Automatically Reducing Visual Clutter”. In: *Smart Graphics*. Ed. by Andreas Butz, Antonio Krüger, and Patrick Olivier. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 77–89. ISBN: 978-3-540-24678-7. DOI: 10.1007/978-3-540-24678-7_8 (cited on pp. 35, 53).

[13] 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. 14, 15, 23).

[14] 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. 28).

[15] 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. 28).

[16] 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. 20).

[17] 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. 24, 25).

[18] 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. 23).

[19] Joseph A. Castellano. *Handbook of Display Technology*. June 1992. ISBN: 978-0-08-091724-5. (Visited on 03/07/2025) (cited on p. 55).

[20] Rebecca A. Champion and Paul A. Warren. "Contrast Effects on Speed Perception for Linear and Radial Motion". In: *Vision Research* 140 (Nov. 2017), pp. 66–72. ISSN: 1878-5646. DOI: 10.1016/j.visres.2017.07.013 (cited on pp. 37, 40, 43).

[21] 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. 22).

[22] 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. 19).

[23] Rune H. B. Christensen. *Ordinal—Regression Models for Ordinal Data*. Manual. 2023 (cited on p. 25).

[24] 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. 15, 23, 47, 49, 51).

[25] 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. 28).

[26] 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. 15).

[27] 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. 22).

[28] *E-Prime*. Psychology Software Tools. 2020 (cited on p. 20).

[29] Gustav Theodor Fechner. "Elements of Psychophysics, 1860". In: *Readings in the History of Psychology*. Century Psychology Series. East Norwalk, CT, US: Appleton-Century-Crofts, 1948, pp. 206–213. DOI: 10.1037/11304-026 (cited on pp. 42, 47).

[30] Stephen Few. "Solutions to the Problem of Over-Plotting in Graphs". In: *Visual Business Intelligence Newsletter* (2008) (cited on p. 35).

[31] 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. 14).

[32] 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 pp. 38, 56).

[33] Arthur P. Ginsburg. “Contrast Sensitivity and Functional Vision”. In: *International Ophthalmology Clinics* 43.2 (2003), p. 5. ISSN: 0020-8167. (Visited on 02/17/2023) (cited on p. 36).

[34] G. Robert Grice, Lyn Canham, and Joseph M. Boroughs. “Forest before Trees? It Depends Where You Look”. In: *Perception & Psychophysics* 33.2 (Mar. 1983), pp. 121–128. ISSN: 1532-5962. DOI: 10.3758/BF03202829. (Visited on 04/17/2023) (cited on pp. 53, 59).

[35] C. G. Healey and J. T. Enns. “Attention and Visual Memory in Visualization and Computer Graphics”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.7 (July 2012), pp. 1170–1188. ISSN: 1077-2626. DOI: 10.1109/TVCG.2011.127. (Visited on 03/13/2025) (cited on pp. 53, 59).

[36] 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. 22).

[37] 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. 27).

[38] Matt-Heun Hong, Jessica K. Witt, and Danielle Albers Szafir. “The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.01 (Jan. 2022), pp. 987–997. ISSN: 1077-2626. DOI: 10.1109/TVCG.2021.3114783. (Visited on 01/28/2025) (cited on pp. 34, 37, 53, 59).

[39] Byron Jaeger. *R2glmm: Computes R Squared for Mixed (Multilevel) Models*. Manual. 2017 (cited on pp. 26, 64).

[40] 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. 30).

[41] Evan Kleiman. *EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data*. Manual. 2021 (cited on pp. 26, 41).

[42] 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. 29).

[43] 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. 27).

[44] 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. 15).

[45] 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 pp. 15, 48).

[46] 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. 27).

[47] Russell V. Lenth. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. Manual. 2024 (cited on pp. 25, 33, 41, 44, 57, 64).

[48] 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. 14).

[49] 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. 25).

[50] 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. 30).

[51] John MacFarlane. *Pandoc*. https://pandoc.org/index.html. (Visited on 10/09/2024) (cited on p. 27).

[52] Justin Matejka, Fraser Anderson, and George Fitzmaurice. “Dynamic Opacity Optimization for Scatter Plots”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 2707–2710. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702585. (Visited on 10/18/2021) (cited on pp. 35, 53).

[53] 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. 28).

[54] 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. 27).

[55] 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 pp. 15, 49, 50).

[56] Stephen R. Midway. “Principles of Effective Data Visualization”. In: *Patterns* 1.9 (Dec. 2020). ISSN: 2666-3899. DOI: 10.1016/j.patter.2020.100141. (Visited on 10/22/2024) (cited on p. 14).

[57] 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. 29).

[58] W. L. Morys-Carter. *ScreenScale*. May 2021 (cited on p. 55).

[59] 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 pp. 26, 64).

[60] Rudolf Netzel, Jenny Vuong, Ulrich Engelke, Seán O’Donoghue, Daniel Weiskopf, and Julian Heinrich. “Comparative Eye-Tracking Evaluation of Scatterplots and Parallel Coordinates”. In: *Visual Informatics* 1.2 (June 2017), pp. 118–131. ISSN: 2468-502X. DOI: 10.1016/j.visinf.2017.11.001. (Visited on 09/08/2022) (cited on p. 48).

[61] 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. 14).

[62] 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. 29).

[63] *OSF*. https://osf.io/. (Visited on 10/10/2024) (cited on p. 30).

[64] 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. 30).

[65] 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. 22).

[66] 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. 20, 38).

[67] 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. 29).

[68] 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. 29, 30).

[69] 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. 27).

[70] Irwin Pollack. "Identification of Visual Correlational Scatterplots". In: *Journal of Experimental Psychology* 59 (1960), pp. 351–360. ISSN: 0022-1015. DOI: 10.1037/h0042245 (cited on p. 49).

[71] 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. 22).

[72] *Prolific.Co*. Prolific. 2024 (cited on pp. 22, 40, 44).

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria, 2024 (cited on p. 23).

[74] 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. 34, 37, 40, 42, 49, 53, 59).

[75] 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 pp. 34, 40, 42, 43, 48, 49, 53, 59).

[76] 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 pp. 15, 49, 51, 59).

[77] Ronald A. Rensink and Gideon Baldridge. "The Perception of Correlation in Scatterplots". In: *Computer Graphics Forum* 29.3 (Aug. 2010), pp. 1203–1210. ISSN: 01677055. DOI: 10.1111/j.1467-8659.2009.01694.x. (Visited on 02/02/2021) (cited on pp. 49, 59).

[78] 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. 29).

[79] 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. 28).

[80] 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. 29).

[81] 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).

[82] 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. 14).

[83] Varshita Sher, Karen G. Bemis, Ilaria Liccardi, and Min Chen. "An Empirical Study on the Reliability of Perceiving Correlation Indices Using Scatterplots". In: *Computer Graphics Forum* 36.3 (2017), pp. 61–72. ISSN: 1467-8659. DOI: 10.1111/cgf.13168. (Visited on 02/02/2021) (cited on pp. 50, 59).

[84] 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. 25).

[85] Maureen Stone and Lyn Bartram. "Alpha, Contrast and the Perception of Visual Metadata". In: *Color and Imaging Conference* 16.355-355 (2008), p. 5 (cited on p. 36).

[86] 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. 15, 23).

[87] 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. 15).

[88] 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. 15).

[89] 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. 16).

[90] Peng Sun, Charles Chubb, Charles E. Wright, and George Sperling. "The Centroid Paradigm: Quantifying Feature-Based Attention in Terms of Attention Filters". In: *Attention, Perception & Psychophysics* 78.2 (Feb. 2016), pp. 474–515. ISSN: 1943-393X. DOI: 10.3758/s13414-015-0978-2 (cited on p. 59).

[91] 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. 29).

[92] *The MIT License*. https://opensource.org/license/mit. (Visited on 10/15/2024) (cited on p. 30).

[93] 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. 28).

[94] Lav R. Varshney and John Z. Sun. "Why Do We Perceive Logarithmically?" In: *Significance* 10.1 (2013), pp. 28–31. ISSN: 1740-9713. DOI: 10.1111/j.1740-9713.2013.00636.x. (Visited on 09/02/2022) (cited on pp. 42, 47).

[95] Cesko C. Voeten. *Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. Manual. 2023 (cited on p. 26).

[96] Paul A. Warren, Laurence T. Maloney, and Michael S. Landy. "Interpolating Sampled Contours in 3-D: Analyses of Variability and Bias". In: *Vision Research* 42.21 (Sept. 2002), pp. 2431–2446. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(02)00266-3. (Visited on 08/22/2023) (cited on p. 59).

[97] Paul A. Warren, Laurence T. Maloney, and Michael S. Landy. "Interpolating Sampled Contours in 3D: Perturbation Analyses". In: *Vision Research* 44.8 (Apr. 2004), pp. 815–832. ISSN: 0042-6989. DOI: 10.1016/j.visres.2003.11.007. (Visited on 08/22/2023) (cited on p. 59).

[98] C. Wehrhahn and G. Westheimer. "How Vernier Acuity Depends on Contrast". In: *Experimental Brain Research* 80.3 (May 1990), pp. 618–620. ISSN: 0014-4819, 1432-1106. DOI: 10.1007/BF00228001. (Visited on 11/02/2022) (cited on pp. 37, 40).

[99] Yating Wei, Honghui Mei, Ying Zhao, Shuyue Zhou, Bingru Lin, Haojing Jiang, and Wei Chen. "Evaluating Perceptual Bias During Geometric Scaling of Scatterplots". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (Jan. 2020), pp. 321–331. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: 10.1109/TVCG.2019.2934208. arXiv: 1908.00403. (Visited on 12/03/2020) (cited on pp. 32, 34).

[100] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. (Visited on 09/14/2022) (cited on pp. 23, 36).

[101] 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. 31).

[102] Yihui Xie. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC, 2015 (cited on p. 27).

[103] 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. 27).

[104] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. "Correlation Judgment and Visualization Features: A Comparative Study". In: *IEEE Transactions on Visualization and Computer Graphics* 25.3 (Mar. 2019), pp. 1474–1488. ISSN: 1941-0506. DOI: 10.1109/TVCG.2018.2810918 (cited on pp. 47, 49, 51).

[105] Silvia Zuffi, Carla Brambilla, Giordano Beretta, and Paolo Scala. "Human Computer Interaction: Legibility and Contrast". In: *14th International Conference on Image Analysis and Processing (ICIAP 2007)*. Sept. 2007, pp. 241–246. DOI: 10.1109/ICIAP.2007.4362786 (cited on p. 36).

Appendices

Appendix A

First appendix

A.1 Section in Appendix