An Empirical Study on Perception of Correlation using Scatter Plots

A thesis presented for the degree of
Masters of Science in Computer Science

Varshita Sher
St. Cross College
University of Oxford
September 2015

Acknowledgments

I would like to express my sincere gratitude to my project supervisor, Professor Min Chen, without whose guidance and motivation, I would not have been able to complete the project. Thanks for all the 'drop-in-any-time' sessions. I could not have asked for a better mentor for my research.

I would also like to thank the rest of the thesis committee members: Karen Bemis, with whom I had brainstorming sessions for hours, figuring out mathematical equations and Ilaria Liccardi, for her insightful comments and suggestions.

My sincere thanks to all my friends and colleagues at Oxford and Department of Computer Science for all the sleepless nights we were working before the deadline. I would like to make a special mention to Viraj and Subhash for the beneficial academic discussions we had and Meenakshi for the ever so often encouragement. I would also like to thank all the participants who took time out of their busy schedule to take part in the study, because of which I was able to draw successful results.

Last, but not the least, I would like to express my deepest gratitude to my family: my parents, my sister and my granny, for providing the much needed emotional and financial support during my time at Oxford, and my life in general. It was their support, inspiration, and sacrifice that helped my pursuit of a high education degree. I am indebted to you all and this is for you!

Abstract

Scatter plots have been in use for over many centuries now but lacks the knowledge of metrics underlying their perception by humans. In this study we empirically assess user performance of estimating correlation in scatter plots for different factors and report whether there is a significant difference and/or relation in the subjective and objective correlation values in relation to different correlation indices, data distribution, symmetry of data enclosure, and number of data points used to plot it. The results suggest that error rates vary in relation to all these factors and condemn the existence of a single linear or non-linear regression pattern to which the human perception of statistical correlation would conform. Even for a set of consistent geometric enclosure such as ellipse itself, the error rate do not conform to a straight line, and thus the possibility of leveraging any perceptual models, such as Weber's law, to evaluate correlation "accuracy" and "precision" is invalid. These finding are significant in that they are new and publishable and falsifies the conclusion of a recent journal paper where the authors might be misled by the accidental patterns resulting from insufficient sampling of a key experimental variable. Lastly, we also establish that as standalone quantities, both human perception as well as the statistical indicator of correlation are unreliable and need to be considered as a 'married couple' which complement each other.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Among all the different kind of statistical graphs available such as parallel coordinate plots, bar charts, bubble plots, scatter plot, network diagrams, and donut charts, Friendly and Denis (2005) emphasizes on the **scatter plots** as the most "versatile, polymorphic, and generally useful invention in the entire history of statistical graphics" (Friendly and Denis, 2005). However, despite their widespread use and importance in science, business, medicine and media, there is a dearth of knowledge regarding their statistical processing, underlying perceptual laws and the way they are perceived by people. This is the primary reason, why their designing process for data presentation and analysis is still largely unscientific (Cleveland and McGill, 1984a). It is important to understand that a wrongly interpreted scatter plot does more harm than good. Thus, the goal should be to generate scatter plots that enable viewers to derive meaningful, correct and unbiased observations from them.

Most of the naive users of scatter plots (sometimes even the experienced ones) are unaware that correlation is not a measure of the appropriateness of the straight-line model. As a result, they base their metrics of correlation estimation on the width of the elliptical shape of the scatter plot. Lewandowsky and Spence (1989) claims that several factors such as unequal variances, outliers, changes in regression slope and point cloud size have a dramatic effect on perception of correlation in scatter plots, but only a few studies provide empirical evidence to support the hypotheses. Bobko and Karren (1979) addressed the issue of misinterpretation of scatter plot by studying via an empirical study the effect of varying range and apparent point cloud shape in scatter plots. Cleveland *et al.* (1982a) established that an observer tends to misinterpret the dataset as highly correlated when the scaling on the horizontal and vertical axes increases. In several experiments (Bobko and Karren (1979); Lauer and Post (1989); Strahan and Hansen (1978)), the effect of varying correlation coefficient

was studied and it was observed that people mostly underestimate correlations over a wide range.

Given all the aforementioned factors that bias user perception while estimating correlation, the basic question that arises is: *"Are scatter plots reliable?". "Is the human perception used to interpret scatter plots, reliable?"*. If yes, then to what extend and under what specific conditions. We aim to answer these questions during the course of the study. We deem this as an important contribution to the statistical and visualization community as the apparent goal of a scatter plot is not just to plot the data points on the coordinate axes as precisely as possible but to see patterns in the data and understand the overall behavior (Cleveland and McGill, 1984b). The research results would be useful to improve basic human judgment required to decode quantitative information from scatter plots which can help improve data display and enhance the accuracy and ability of observers to visually decode the data more efficiently and accurately.

1.2 Objective

The goal of this project is to **study the human perception of correlation in scatter plots**. We will focus our study on the basic visualization task of estimating correlation in bivariate data. We also **study whether the "precision" and "accuracy" for the perception of correlation in scatter plots are systematically linked via Weber's law**. Here "precision", refers to ability to detect the difference between two correlation value, even if they are unaware of the actual numerical correlation value and "accuracy" refers to the ability to make an estimate of the value of correlation, which as close as possible to the actual correlation value.

We use an empirical approach to provide concrete evidence for the findings of our study. The empirical study indicates whether the observed factor has any enhancing or deteriorating impact on the accuracy of the user estimate of correlation. Thus, we summarize the objective of our study as follows:

1. To study factors which humans use while examining a scatter plot which may potentially impact user perception when presented with scatter plots to estimate correlation.

2. To formulate hypothesis for the empirical study, indicate measurement metrics and identify the statistical analysis to analyze the results.

3. To design task and stimuli pertaining to each of the hypothesis and identify experiment variables for each task.

4. To design software to conduct the experimental study and gather user response.

5. To analyze the empirical study results.

As far as our knowledge goes, none of the previous studies have been aimed specifically at studying the design variation in point cloud shape in scatter plots. Our study researches the "science" of scatter plots through human graphical perception. Our approach is both theoretical and experimental and would provide concrete evidence for appropriate design and usage of scatter plots so as to achieve highest level of visual understanding and prove beneficial for the information visualization community.

1.3 Structure

The entire dissertation in laid down in seven chapters, including this introductory chapter explaining the motivation behind the study, the objective of our research and the outlining structure of the dissertation.

Chapter 2 will give some background information regarding the general statistical concepts of scatter plot and correlation and methodology implemented in an empirical study and how the subsequent analysis of the result is performed.

Chapter 3 will focus on the related research previously done on these topics.

Chapter 4 will cover the methodology used in the study and formulate formal hypothesis. It explains the tasks performed as part of the study and identify the independent, dependent and control variables for each.

Chapter 5 will discuss the designing of the user study which includes the formulation of the tasks, the stimuli (scatter plots) and the software.

Chapter 6 will describe the implementation of the all the study elements including stimulus and software. It also explains the actual implementation of the software.

Chapter 7 will analyze the participant result and validate the hypothesis formulated at start of the study along with summarizing user performance in each task and inferring conclusions.

Chapter 8 will summarize the result of the research in a nutshell and provide directions for future work.

The appendices at the end further facilitate the dissertation by providing all stimuli used in the study, the data tables containing measurements for the creation of scatter plots, the introductory PowerPoint presentation to introduce participants to the empirical study and basics of scatter plots and correlation, the feedback survey form presented to collect subjective data and schedule of our project.

Chapter 2

Background

This chapter gives a brief introduction to some of the statistical concepts and empirical methodologies, used as part of our research in general. We assume pan-sector audience at the receiving end of this dissertation and thus dedicate a chapter to explain the underlying foundational theory of this research.

We begin with explaining in Section 2.1 what correlation means in statistics, some of its applications in different fields for problem solving and the types of correlation coefficients along with some of its properties. We then move on to briefly introduce scatter plots in Section 2.2 and their association with correlation. Section 2.3 focuses on the "perceptual law" namely, Weber's law that has been previously used to lay down the foundation for explaining suitability of a visualization design to a particular task category and the associated concept of Just Noticeable Difference (JND). Lastly, we enlist the steps to organize our approach towards an empirical study in Section 2.4.

2.1 Correlation

Correlation, in simple terms means the amount of relationship or connection between two or more things. Mathematically speaking, correlation (henceforth denoted by R), is the degree to which two variables are associated. It gives an estimate of how change in one variable's value is associated to a change in the other. For example, relation between height and weight of a person or between a person's smoking habits and probability of him being affected by lung cancer.

2.1.1 Applications of Correlation

Correlation has been useful in the past for "identifying pattern trends and make future predictions based on it" (Kanjanabose, 2014). Mowry and Luk (1997) introduced a technique called "correlation profiling" to predict data cache misses more accurately

where "Correlation profiling" involves investigating the relation between cache misses and related information such as recent control-flow path and cache outcomes of previous references. (Alexiadis *et al.*, 1999) developed an Artificial Neural Network (ANN) that reliably forecasts wind speed for several hours to efficiently accommodate the wind generation based on spatial correlation models. They used "persistent forecasting" to observe correlation between consecutive wind-speeds at different sites, pressure profiles and terrain shapes. More recently, correlation has also been used in clinical studies for exemplifying the data from a sample of women attending their first antenatal clinic visit which studies the strength of the relationship between maternal age and parity (Mukaka, 2012).

Recent reports have established that stock market volatility is time varying and even exhibits positive serial correlation (Maheshchandra, 2012). Correlation has also found usefulness in building a stock portfolio by observing the impact of stock with respect to market, which has displayed synchronized movement (Ding *et al.*, 1993). Ding *et al.* (1993) investigated a 'long-memory' property of the stock market returns series and reported that there is substantially more correlation between absolute returns than returns themselves. 'Long memory' dynamics are important pointers for identifying presence of nonlinear relations in conditional mean and variance of financial time series.

Thus, in the past, the applications of correlation have been manifold and are continually expanding to almost every sector. We next explain the mathematical formulas useful for calculating the correlation values in statistics.

2.1.2 Types of Correlation Coefficient

The correlation coefficient is the measure of the degree (strength) and direction (positive or negative) of the correlation between two or more variables. It is expressed using the variable r, where r lies between -1 and $+1$, inclusive. The closer r is to these two values, the less scattered the points are and stronger is the relationship. When $r < 0$, the data has a negative correlation, and when $r > 0$, the data has a positive correlation. The data set is said to be *perfectly positively correlated* when $r = 1$ and *perfectly negatively correlated* when $r = -1$.

There are two different formulas to be used while calculating correlation coefficient. For normal distributions, we use the Pearson Product-moment Coefficient (r_p) whereas we use the Spearman Rank-order Coefficient (r_s), when the data is ranked (1^{st}, 2^{nd}, 3^{rd}, etc.).

2.1.2.1 Pearson coefficient of Correlation (r_p)

The Pearson correlation coefficient was introduced by Galton in 1877 (Galton, 1888; Stanton, 2001) and developed later by Pearson (Pearson, 1904). It measures the degree of *linear relationship* between two variables which are normally distributed (Altman, 1990). A relationship is linear when a change in one variable is associated with a proportional change in the other variable. For variables X and Y,

$$r_p = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)(S_X S_Y)}$$

where,

$\overline{X}$ = sample mean of the first variable i.e. X

$\overline{Y}$ = sample mean of the second variable i.e. Y

S_X = standard deviation for the first variable i.e. X

S_Y = standard deviation for the second variable i.e. Y

n = number of data points in the sample

2.1.2.2 Spearman's coefficient of Correlation (r_s)

The Spearman correlation coefficient evaluates the *monotonic relationship* between two continuous or ordinal variables which is highly skewed (Altman, 1990; Zou *et al.*, 2003). In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data. For variables X and Y,

$$r_p = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where,

d_i = the difference between ranks of corresponding values X_i and Y_i

n = number of data points in the sample

2.1.3 Properties of Correlation coefficient

Correlation is a statistical method used to assess a possible linear association between two continuous variables. It is simple both to calculate and to interpret. However, misuse of correlation is so common among researchers that some statisticians have wished that the method had never been devised at all (Altman, 1990). We therefore enumerate some of the properties regarding it's correct usage and application.

1. The correlation coefficient is a measure of the strength of the linear trend relative to the variability of the data around that trend. Thus it depends on the amount of linearity as well as the amount by which it varies.

2. Correlation does not imply causation. There could be several possible explanations for a strong correlation observed between two variables, A and B,

   - A influences B
   - B influences A
   - A and B are influenced by one or more additional variables
   - The relationship observed between A and B was a chance error

3. The correlation coefficient is a dimensionless quantity.

4. The linear correlation estimators are used for estimating the standardized shape parameter of an elliptical distribution which generally has the well known normal distribution. But rarely ever, the data is normally distributed and hence correlation coefficient has a very bad performance for heavier tailed or contaminated data.

5. It measures the strength and direction of **linear relationship** between two variables **only**. If a non-linear relationship exists, it may be inadequately described by r or possibly even undetected. Hence, in addition to a numerical correlation coefficient, there is a need for a supporting scatter plot visualization, which is explained briefly in the next section.

2.2 Scatter Plots

According to Cleveland, "scatter plot conveys an understanding of the relationship between two variables which may be for the purpose of data exploration or to assess a model's fit to the data, two general functions of statistical graphs" (originally by (Cleveland and McGill, 1984a) cited in (Lauer and Post, 1989)). Friendly and Denis (2005) explains a scatter plots as a "plot of two variables, x and y, measured independently to produce bivariate pairs (x_i, y_i), and displayed as individual points on a coordinate grid typically defined by horizontal and vertical axes, where there is no necessary functional relation between x and y".

In general, scatter plots (also called scatter diagrams) are used to visually investigate the possible relationship between two variables. Even though they lack some of the flexibility and visual expressiveness rendered by newer multidimensional visualization techniques, they are widely used due to their simplicity, familiarity and visual clarity. They complement correlation calculations in that it aims at displaying pictorially all the data and proximity between data points rather than just a statistical numeric value (Li *et al.*, 2010).

The scatter plots were originally designed to process bivariate data only but recent developments show that these have been developed to display three or more variables using various packages (Ligges and Mächler, 2002), 3D plots (Elmqvist *et al.*, 2008) and brushing techniques (Becker and Cleveland, 1987). Becker and Cleveland (1987) uses 4 main brushing operations (highlight, shadow highlight, delete and label) to interactively view multidimensional data. It displays the pairwise scatter plots of the variable in a rectangular array called scatter plot matrix such that the brushing techniques on one of the scatter plot simultaneously induces the effect on all scatter plots. Elmqvist *et al.* (2008) uses a matrix of scatter plots and transitions between them in 3D space to interactively navigate through all the possible configurations. Ligges and Mächler (2002) implements an *R* package for the visualization of multivariate data using parallel projections, utilizing some pre-existing functionalities of *R*, which is a software tool that provides an environment for statistical computing.

2.3 Just Noticeable Difference

The concept of just noticeable difference (JND) is the result of seminal works in the area of classical psychophysics conducted in the mid-nineteenth century Stern and Johnson (2010). Several definitions of JND have been provided over the years. Rensink and Baldridge (2010) describes the JND as the *"difference in properties between two side-by-side stimuli (e.g. squares of differing brightness) that can be discriminated* 75% *of the time"*. Harrison *et al.* (2014) describes it as the *"quantity by which a given stimulus must increase or decrease before humans can reliably detect changes"*. In simpler terms, the JND is the minimum amount of adjustments that need to be done in a stimulus in order to produce a noticeable variation in sensory experience. For example, JND is be the minimum amount by which a person must raise his voice in order to be audible in a noisy room.

It was Ernst Weber (Weber, 1978), who first discovered that this minimum amount is lawfully related to the initial stimulus magnitude and coined the term, Weber's law as explained in the next section.

2.4 Weber's law

Weber's law is a psychological law quantifying the perception of change in a given stimulus. It states that the amount by which a physical stimulus must be increased (or decreased) in order for the difference to be detected by an observer is a constant fraction of the intensity of the original stimulus (Lewandowsky and Spence, 1989). According to it, the JND in a physical property is a fixed proportion of its magnitude i.e. $\Delta I = KI$ where is JND corresponding to reference stimulus I and K is the Weber

fraction (Lewandowsky and Spence, 1989). For example, in a quiet room it is easy to hear someone whisper but for a person to be audible in a noisy room he has to shout.

A small Weber fraction indicates a fairly high sensitivity to JND variations, while a larger Weber fraction is rather poor and indicates lower sensitivity to variations (Halberda). This in-depth knowledge of JND allows graphic designers to design better interface and facilitate interaction as users can fully grasp differences in objects, shapes, texts, etc influencing their cognition abilities.

Smeets and Brenner (2008) emphasized that Weber's law can hold only for physical properties that have a magnitude i.e. those that start at an absolute zero and cannot be negative. Hence Weber's law holds for perception of size, weight and distance but not for shape, orientation or position.

According to various statistical and psychophysical books (Cornsweet, 2012; Stevens, 1975), Weber's law expresses the equation for a straight line ($y = mx + b$), with the slope m being the Weber Fraction, K and the y-intercept b being zero . Likewise, I plays the part of x and ΔI plays the part of y. Thus, if Weber's law holds, we expect the data to graph as a straight line. On the other hand, in case of failure of Weber's law, we obtain a curve with polynomial, logarithmic or exponential growth. Studies conducted by Bizo *et al.* (2006) demonstrated that a U-shaped pattern was observed for Weber's fraction rather than a linear regression line when studying Weber's law in temporal production and categorization tasks (Bizo *et al.*, 2006).

2.5 Empirical Study

The word empirical means information gained by observation, or experiment. An empirical study is a commonly used method to compare user performance in various tasks. It helps answer the proposed research questions based on concrete evidence from observed and measured formula rather than from theory or belief.

The basic steps involved in an empirical study is summarized in the flowchart in Figure 2.1. We begin by formulating the research questions we want to study (Kanjanabose, 2014). We then proceed to design the hypotheses for the experiment and the relevant dependent, independent and control variables such that the confounding effect is minimized (Kanjanabose, 2014). Once the designing is done, an experiment is conducted to collect the user data which is then statistically analyzed to understand the research outcomes and subsequently evaluate the hypotheses (Kanjanabose, 2014).

Following subsections will explain these processes in detail, starting from formulating research question and hypotheses, identifying variables in the experiment, minimizing the confounding effect and performing statistical analyses.

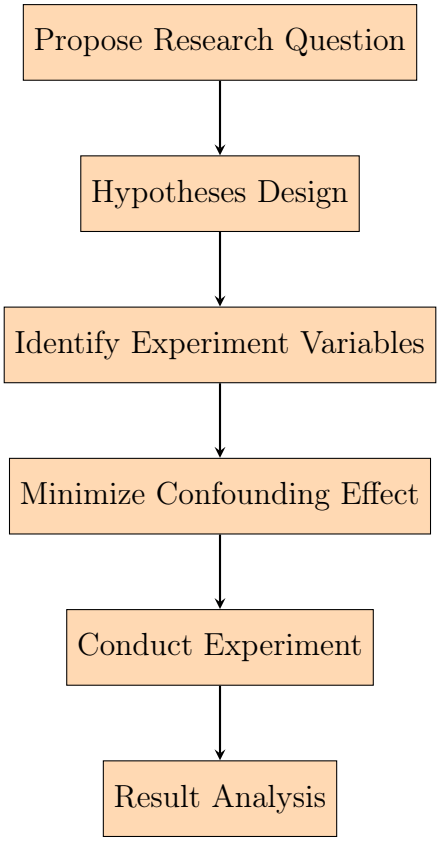


Figure 2.1: Workflow of an Empirical Study

2.5.1 Research Question and Hypotheses

Kanjanabose (2014) explained that the research question is the initial step in a research project and is explicitly mentioned at start of the study to familiarize the readers with the objective of the study. It is a non-biased and precise question which helps you keep focused on the area of enquiry rather than getting side-tracked (Springett and Campbell). Lastly, it is a good practice to define any key terms in the research project or research question upfront so that everyone has a shared understanding (Springett and Campbell).

We next develop the hypotheses relevant to the research question (Kanjanabose, 2014). While the research question is broad and includes all the variables you want your study to consider, the hypothesis is a statement that specifies relationship you expect to find from the examination of these variables and is testable and falsifiable (Rubin, 1975).

2.5.2 Variables in Experiments

A typical experiment consists of three main variables which are likely to vary or remain constant throughout the course of experiment, namely: *independent variables, dependent variables* and *control variables*. In addition to these, we introduce another

category of variables, exclusively for our study called *task variables*. We explain each of these below.

Independent variables and its levels

These are the factors that are studied as part of the experiment to see whether they influence other factors or not. For example, the temperature on a given day (in $^{\circ}C$) is an independent variable.

In an experiment, the independent variable is varied or manipulated by the experimenter so that its effect on other variables can be measured. A single condition or treatment of an independent variable is called a **level**. For example, if the independent variable was temperature (in $^{\circ}C$), each of the temperatures on different days used would be a level i.e $10^{\circ}C$, $20^{\circ}C$ and $30^{\circ}C$ would all be levels of the independent variable.

An independent variable must have at least two levels (variants) in order for comparison to take place between them (Fraenkel *et al.*, 1993; Kanjanabose, 2014; Rubin, 2012) and it is generally plotted on the x-axis on the graph.

Dependent Variable

The dependent variables are the factors influenced by the independent variables. They are compared across different levels of independent variables during an experiment which tells whether independent variable influences a dependent variable or not. It's value depends upon the value of the independent variable and is generally plotted on the y-axis on the graph. For example, the sale of ice-creams at an ice-cream parlour on a particular day is a dependent variable, depending on the temperature on that day.

In our study, while evaluating the effect of different scatter plot design factors on perception of correlation, the factors will be the independent variables and the user perception is the dependent variable.

Control Variables

These are the variables which can affect our results and influence the measured factors, if we do not control them. The researcher wants to ensure that it is the manipulation of the independent variable that has caused the changes in the dependent variable and not any *extraneous* or *confounding variables*. Hence, (Kanjanabose, 2014) emphasized that these variable must be held constant to avoid misleading results due to **confounding effect**.

In the aforementioned example of studying the relation between the temperature and ice cream sale, we must keep the reference Ice-cream parlour constant while collecting observations on different days.

2.5.3 Confounding Effect

Confounding is defined as a situation in which the effect of the two processes cannot be separated. It may produce spurious results, leading us to believe the existence of a valid statistical association when one does not exist or alternatively the absence of an association when one is truly present. The within-subject experimental designs where each participant is given every possible treatment (subjected to all independent levels), has many advantages but suffers from the most common confounding effect, namely **carryover effect**. In a carryover effect, one treatment affects subsequent treatments and thus the results obtained are biased. It has two kinds namely, **sequence effect** and **order effect**.

A sequence effect is related to effects from a preceding treatment whereas an order effect is associated more with the position of a treatment condition in a sequence of treatment conditions. There isn't much difference between the two in context of a two-treatment experiment where order and sequence are tightly bound together as the second treatment can follow only the first treatment and can only have the order 2. However, when participants are subjected to multiple treatments, we can isolate the effects of the two. For example, in both sequences "A B C" and "C B A", the order of B is 2 but its preceded by 'A' in the former setting and 'C' in the latter.

The carryover effects impact user performance because the earlier trials provide a learning curve to the participants due to which they gain some amount of experience and perform better. It may also be the case that their performance may deteriorate as continued trials bring fatigue and boredom (Kanjanabose, 2014).

2.5.3.1 Controlling confounding effect at the design stage

In case of relatively fewer experimental conditions, Salkind (2010) emphasizes on using complete counterbalancing of experimental conditions such that each subject is exposed to every possible combination in one order and then again in reverse order. Thus, the total number of sequences needed, for each task is n where n is the number of possible treatment conditions (levels of the independent variable).

Salkind (2010) suggests another strategy if the counterbalancing within subjects is not feasible due to the need for examining numerous conditions. It includes counterbalancing conditions between subjects using a *balanced Latin square* where a limited set of permutations is used. Each treatment is preceded and followed, equally often by all the other treatments and administered once in each ordinal position (first, second, third).

For example, for $n = 4$ where n is the number of conditions, instead of requiring 4 possible sequences, we follow the Latin square arrangement of treatments, specified in the following table for testing all the four levels of the independent variable. Here,

each condition is preceded by every other condition equal number of times.

Sequence 1	A	B	D	C
Sequence 2	B	C	A	D
Sequence 3	C	D	B	A
Sequence 4	D	A	C	B

Apart from the aforementioned techniques, there are various others ways to exclude or control confounding variables. For the design stage, *randomization, restriction*, and *matching* of the dependent variables is prescribed by (Aschengrau and Seage, 2008; Jepsen *et al.*, 2004). This means that all study participants or samples should be randomly selected from a pre-set range to reduce the possibility of chance.

2.5.3.2 Controlling confounding effect at the analysis stage

Most often, the experimental designs are premature, impractical or impossible at the design stage, and hence the use of statistical methods such as **standardization, stratification** and **multivariate models** is emphasized upon to adjust potentially confounding effects at the analysis stage (Aschengrau and Seage, 2008; Pourhoseingholi *et al.*, 2012).

Out of these three, when the sample size is large, such as in our case, multivariate models are the most efficient choice (Afifi *et al.*, 2011). It includes techniques such as linear regression, ANOVA analysis, Friedman Test and post-hoc test, some of which we use in our study, as explained in the next section.

2.5.4 Statistical Analysis

Once the data from the user study has been collected, it is scrutinized to draw valid inferences from them. In most researches conducted, descriptive and inferential statistics are the two main types of statistical analyses performed to analyze the results and draw conclusions (Kanjanabose, 2014; Kolaczyk and Csárdi, 2014; Vergura *et al.*, 2009). Descriptive statistics allows us to summarize our data in meaningful ways so that the patterns and trends are evident whereas inferential statistics determine how likely it is that the results based on a particular sample are similar to results that would have been obtained for an entire population (Fraenkel *et al.*, 1993).

Descriptive statistics enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. It primarily consists of *measures of central tendency* and *variability* (Kanjanabose, 2014).

Inferential statistics primarily constitutes the hypotheses testing (Kanjanabose, 2014). We begin by formulating the null hypothesis ($H_\circ$) which is the assumption that

there is no significant difference amongst the independent variable levels. In a mathematical formulation of the null hypothesis there will typically be an equal sign. The alternative hypothesis (H_1) states that there is at least one level that is significantly different from the others. In a mathematical formulation of the alternative hypothesis there will typically be an inequality ($<, >, \leq, \geq$) , or not equal to symbol ($\neq$). Rejecting the null hypothesis from an inferential analysis means the difference in the results did not occur by random chance and a real effect of the independent variable exists which can be generalized. However, failing to reject a null hypothesis can be the outcome of two conditions, *a*) ($H_\circ$) is actually true or *b*) ($H_\circ$) is false, but we have not collected enough data to provide sufficient evidence against it (Kanjanabose, 2014).

Because the hypothesis testing is based on probabilities, there is always a chance of drawing an incorrect conclusion. While performing a hypothesis testing, two types of error are possible, 1) the researcher rejects a null hypothesis when it is true, causing Type I error or 2) fails to reject a null hypothesis when it is actually false, causing Type II error. The probability of committing a Type I error is called the *significance level*, and is often denoted by α. The significance level is usually set at 0.05 or 0.01. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists. Thus, for our study we agree upon $\alpha = 0.05$, which indicates we are willing to accept a 5% chance that we are wrong when we reject the null hypothesis.

To reject the null hypothesis, the obtained value from the hypothesis testing called the p-value must be less than the significance level (α). A p-value is described as the probability of obtaining the observed results or a more extreme result, if the null hypothesis is true (Cumming, 2008; Lew, 2012). Mathematically speaking, it is expressed as: p-value=Pr[observed or more extreme data$|H_\circ$]

If the p-value < 0.05, we reject the null hypothesis, otherwise we fail to reject the null hypothesis. Furthermore, rejecting the null hypothesis means that there is a statistically significant difference between the means of each group of independent variables and failing to reject it means they are all the same.

2.5.4.1 Friedman Test

Friedman test is a non parametric test comparing differences between groups when the dependent variable being measured is ordinal (in our case, the accuracy of user performance). It doesn't specifically require the data to fit a normal distribution and works well with skewed distribution. Since, we work upon participant data which can be quite unpredictable, depending on each individuals capabilities, Friedman test prove to be the most appropriate choice for comparison of medians of more than two groups to detect presence of a significant difference between them.

The null hypothesis (H_0) states that the medians of the c groups are equal whereas the alternative hypothesis (H_1) states that at least two medians are significantly different. If the Friedman test shows p-value < 0.05, we reject the null hypothesis and report that a significant difference exist. Otherwise we fail to reject the null hypothesis.

The Friedman test first ranks the values in each row (block) from lowest to highest, resolving ties by assign them the mean of the ranks that they would otherwise have been assigned. Each row is ranked separately. It then sums the ranks in each column (group). If the sums are very different, the p-value will be small and thus we can safely reject the null hypothesis along with the idea that all of the differences between columns are due to random sampling, and conclude instead that at least one of the groups differs from the rest.

We define below the Friedman test statistic F_R (or **chi-square** statistic) for the Friedman rank test.

$$F_R = \frac{12}{rc(c+1)} \sum_{j=1}^{c} R_j^2 - 3r(c+1)$$

where

R_j^2 = square of the total of the ranks for group j $(j = 1, 2, .., c)$

r = number of blocks

c = number of groups

The p-value is obtained from the test statistic F_R, by approximating it using the chi-square distribution with degrees of freedom $c - 1$, where c is the number of groups.

Lastly, we report the results of a Friedman Test using the test statistic (χ^2) value *(chi-square)*, degree of freedom **(*df*)** and the significance level *(p-value)* as : ($\chi^2(df)$ = chi-square , p = p-value). The *df* is equal to $c - 1$ where c is the number of groups being tested.

2.5.4.2 Wilcoxon Signed-Rank Test

Friedman test is an omnibus test that just indicates presence of overall differences. Therefore, we use the Wilcoxon test which is a post-hoc test aimed at identifying the exact pairs of groups that significantly differ from each other.

The null and alternate hypotheses of Wilcoxon test are similar to that of the Friedman test. Wilcoxon Signed-Ranks statistic follows the z distribution when the sample size is large. It first calculates the difference between the scores of the two groups and later ranks them irrespective of the sign of the difference (ignoring differences of 0). Next, it adds up the rankings of both the positive scores and the negative scores and the smaller of the two, becomes the Wilcoxon test statistic W. Lastly, W is used to calculate the z- statistic using the formula :

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

where,

n = the number of differences (omitting 0 differences).

The z-statistic is then used to read the critical values for the 2 tailed test at $\alpha = 0.05$ level of significance, called the 2-tailed asymptotic significant value or in general, the p-value. As in the case of Friedman test, if the p-value < 0.05, we reject the null hypothesis else fail to reject it. Finally, we report the results from the Wilcoxon test using the **z-statistic** and the **p-value**: *(Z = z-statistic, p = p-value).*

Chapter 3

Related Research

This chapter gives an overview of the previous related research done in the field of visualization aiming mainly at studying perception of graphs in general and scatter plots in particular. We discuss the relevant theoretical literature and lastly we point out the areas which are in need of more empirical work for concrete results.

3.1 Graphical Perception

Graphs have played an essential role of conveying information in a pictorial manner for about 200 years and humans tend to base their decision making on inferential judgment from them (Bobko and Karren, 1979). However, only a few studies have been aimed solely at understanding the psychophysical approach deployed by humans for understanding the data presented in these graphs and the various design factors that tend to make an impact on their perception. Scatter plots, in particular, have been faintly studied, despite their utmost importance, so much so that have even been used as evidence in equal employment opportunity court cases, where its believed that few of the participants were unable to perceive the correlation in them correctly (Bobko and Karren, 1979), resulting in biased legal judgments.

Although scatter plots are routinely used and their advantages are multifaceted, the potential to derive benefits from it depends largely on the observer's ability to perceive and interpret the graph correctly (Lewandowsky and Spence, 1989). As a result, they can be often misleading when using them to judge correlation (Li *et al.*, 2010; Loh, 1987), a task identified as one of the "10 low-level analysis task" that are important within Information Visualization field (Li *et al.*, 2010).

Most often, the true relationship between variables are misinterpreted because human perception of correlation is based on assumption of linearity and homoscedasticity (Doherty *et al.*, 2007). We need to be vary of it since a wrongly interpreted scatter plot does more harm than good as it becomes a problem to be solved rather

than an aid to understanding (Kosslyn, 1985; Lauer and Post, 1989). (Cleveland and McGill, 1984b) enumerated, what they call the *set of elementary perceptual tasks*, or in simpler terms, judgments carried out by an observer to extract quantitative information from graphs and later ranked them in the order of how accurately people perform them. They enumerated six basic judgments namely, position along a common scale, position along identical but non-aligned scale, length, angle, slope and area and came to the conclusion that the two position judgments are the most accurate, followed by length judgments. Angle and slope judgments are third, and least accuracy is observed for area judgments. Such concrete results are missing in context of scatter plots.

Apart from these, there is relatively less knowledge about the metrics that humans deploy which aids better visual processing of the scatter plots (Lewandowsky and Spence, 1989). Hence, it's essential to decipher the "perceptual, psychophysical and cognitive processes" that are invoked during the examination of a scatter plot for a visualization task (Doherty *et al.*, 2007; Lewandowsky and Spence, 1989). Doherty *et al.* (2007) claims that such a study that compares the intuitive estimate with the actual value of the numeric correlation is essential as it helps in determining "the relative distances between representations in psychological space". In presence of it, we can reason about the processing methodology of people, of all statistical expertise levels, when perceiving data presented in scatter plots.

Previously, two main methodologies have been used for investigating presence of correlation in scatter plots - *theory driven* where people tend to use their preconceived knowledge and draw false conclusions by mostly overestimating correlation and *data driven* where people have no prior beliefs or domain knowledge and hence their estimates of correlation are mostly underestimated (Meyer and Shinar, 1992). Previous researches have been focused mainly on theory-based estimates and data-driven estimates haven't been explored much, despite the fact that graphical displays are becoming an integral part of decision-making processes as numerous graphic software packages are being introduced everyday. It is therefore essential to study the cognitive aspect of intuitive assessment of correlation in scatter plots.

The most common mistake humans make while reading a scatter plot is interpreting the linear degree of the point cloud therein, as a representative of the absolute value of correlation (Li *et al.*, 2010; Loh, 1987), which may lead to erroneous observation, unreliable analyses and faulty conclusions. The correlation is a measure of the degree of linear association between two normally distributed variables only rather than the degree to which the point cloud is linear. Figure 3.1 illustrates rotation of a point cloud which yields different correlation values while the linear degree of point cloud obviously remains the same. In a similar experiment, Cleveland *et al.* (1982b) observed two ratios to determine whether they hold any significant relation to judg-

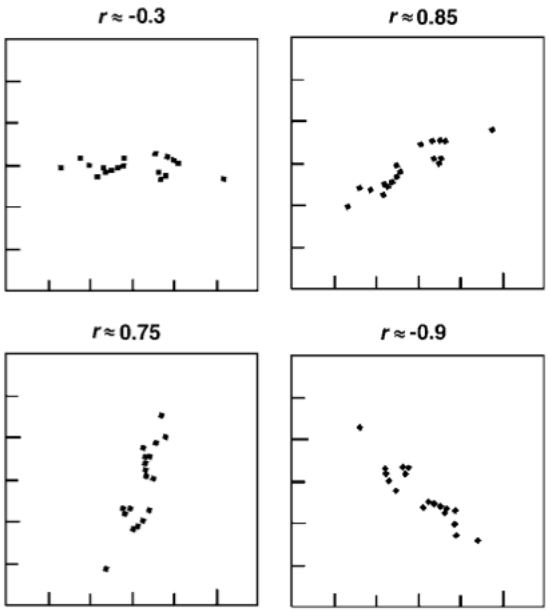


Figure 3.1: Effect on correlation of rotating a point cloud in a scatter plot (Li *et al.*, 2010)

ment of correlation. They firstly examined the ratio of the size of the elliptical point cloud to the size of the circumscribed rectangle and then the ratio of the minor axis to the major axis of the point cloud. However, they found that neither of them could be sufficiently used to describe the judgment of correlation in the data, which means that the human perception is mostly unpredictable and hence unreliable.

It has been established in previous studies, that humans, irrespective of their statistical expertise level, tend to underestimate correlation. Meyer and Shinar (1992) studied the relationship between intuitive estimate of correlation and the actual correlation coefficient values and the varying effect of various display characteristics of scatter plots on user estimates by recruiting participants of varying statistical training (department faculty vs. undergraduate students). They concluded that the statistical knowledge is irrelevant with respect to perception as the visual properties of display influence the "cognitive structures created by formal statistical training" (Meyer and Shinar, 1992).

Strahan and Hansen (1978) conducted an experiment to observe user's estimate of correlation for bivariate normal scatter plots with equal variance of the two variables, plotted with 200 data points and correlation values ranging from 0.010 to 0.995. The results revealed that users, although statistically sophisticated, underestimated correlation values, with greater deviations in the middle as compared to the extremes of 0 and 1. An year later, Bobko and Karren (1979) conducted another study using questionnaires to investigate the perception of correlation coefficient from scatter plots. Their results were consistent with those of Strahan and Hansen (1978). In addition,

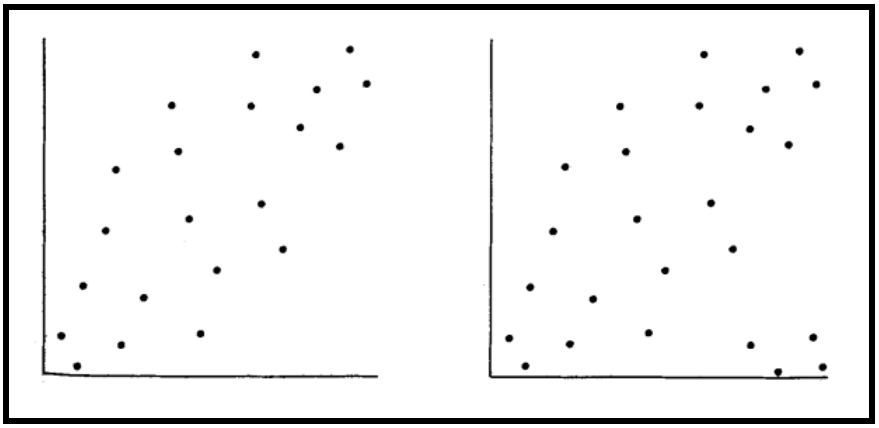


Figure 3.2: Two scatter plots with $r = 0.72$ and $r = 0.27$ respectively. They differ only with respect to four outlier points in lower right corner of the second display. (Bobko and Karren, 1979)

they narrowed the source of the underestimation and claimed it was more pronounced in the range $0.2 < |r| < 0.6$.

Among all the prominent factors for correlation underestimation, the majority of biased results can be accredited to people's sensitivity to "contaminated" scatter plots (a term coined by Konarski (2005)), where "contamination" refers to presence of outliers in data distorting the cloud shape. Although, Lew claims humans to be better judges of correlation coefficient, in presence of outliers, compared to some robust numerical estimators (Lewandowsky and Spence, 1989), previous studies depict that the amount of distortion in the scatter plot from the outliers is underestimated by human's intuitive estimate of correlation. Their estimate isn't as biased as should be from the presence of outliers. Bobko and Karren (1979) studied the effect of outliers by generating four additional points in the lower right corner of a perfectly elliptical scatter plot (see Figure 3.2) and concluded that users underestimated the attenuating effects of outliers. Their results were consistent with those of Bobko and Karren (1979); Meyer and Shinar (1992) and revealed that people tend to underestimate values, ignoring the smaller distribution of outliers.

Lauer and Post (1989) conducted an empirical study of scatter plots where participants estimated correlation using a mouse to select a value on a number line that was displayed below each scatter plot. The user estimates were judged on basis of influence by a number of factors including the objective r, the variance of X, the variance of Y, the regression slope, dispersion of point cloud, density (number of data points) of the scatter plot and even size of the CRT screen. Their results demonstrated that all of these aforementioned factors were found to significantly affect subjects' estimates of correlation. A significant underestimation for r was observed with the mean estimate being a sharply positively accelerated function of r.

The dispersion of data cloud is perhaps the most influential factor affecting in-

tuitive estimation of correlation (Doherty *et al.*, 2007). Most people, or "intuitive statisticians" as Meyer refers to them (Meyer and Shinar, 1992), base their estimate of correlation on the 'thickness' or 'thinness' of the ellipse present in the scatter plot. This is a very dangerous assumption as hardly ever do we find empirical data of such nature which fits an elliptical shape of absolute normal distribution.

As far as our knowledge goes, the study of perception based on the cloud shape of scatter plot has been performed only by Bobko and Karren (1979) and Meyer and Shinar (1992). Bobko and Karren (1979) studied the "twisted pear" shape (see Figure 3.3) which Fisher (1959) claimed to represent patterns that are "characteristic of prediction problem" across wide content range area (Bobko and Karren, 1979). Meyer and Shinar (1992) investigated the "trumpet-shaped" dispersion (see Figure 3.4) in which the variance between data points increased at higher levels of first variable (x) and found that users chose to concentrate on majority of the x level where the data points were concentrated around the regression line and ignore higher dispersions at upper end of the x-axis. Their study considered only three correlation values: $r^2 = 0.4, 0.65$ and 0.9 with the increase in variance of x only. Due to this, their study was inadequate to figure out whether any of the negative correlations has an equal amount of impact on perception of due to variation in variance of variables. However, there are many variation other than the aforementioned ones which need to be studied and quantified with respect to the error rate in user perception and is a peculiarly interesting topic of research. In our research, we study the effect of few other scatter plot shapes that are remarkably different from the standard elliptical fit of data.(see Figure 3.5), to study whether participant perception is reliable enough to successfully figure out the correlation value from them.

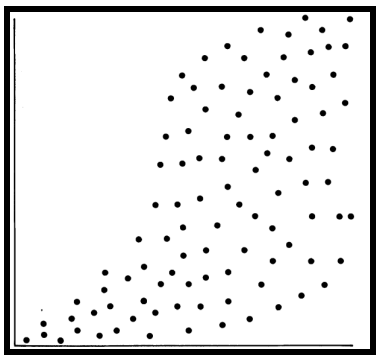


Figure 3.3: Scatter plot in shape of twisted pear ($r = 0.6$) (Bobko and Karren, 1979)

It seems clear that a myriad of factors such as scaling, number of points, display device affect the appearance of the scatter plot and subsequently influence correlation estimation judgments. Only a few of them have been pursued and explored in the past, as the study design of the research is considered prohibitively costly in terms of time and effort (Doherty *et al.*, 2007).

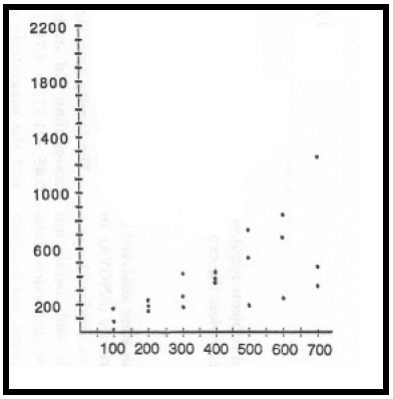


Figure 3.4: Trumpet-shaped dispersion of data points ($r = 0.65$) (Meyer and Shinar, 1992)

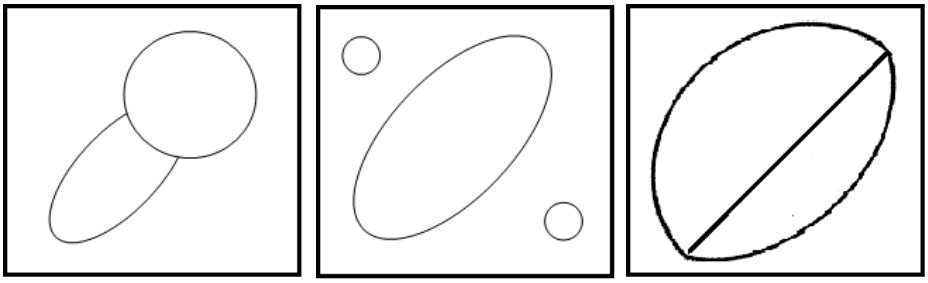


Figure 3.5: Different cloud shapes of scatter plots investigated in our study

Our systematically planned study, aims to investigate the impact of various statistical and perceptual properties of scatter plots on correlation estimation. We believe that proficient comprehension of some of these properties may help graphic designers choose appropriate scaling and formats to minimize misconceptions about the relative strength of association within data and improve credibility of the information conveyed by the scatter plot. We could utilize the findings from this study to design scatter plots that provide impressions of displayed data that are easily understood by almost everyone.

Chapter 4

Methodology

In this chapter, we will first explain each task in detail and then formulate 6 hypothesis for each of the 6 task categories namely, *JND, Weber, Distribution, Density, Progressive Symmetry* and *Reflective Asymmetry*. To evaluate these hypothesis, we will design and conduct a controlled experiment with relevant independent and dependent variables. The result from the experiment will then be analyzed using descriptive and inferential statistics including Friedman and Wilcoxon signed rank test, explained in Section 2.5.4 to provide statistical evidence in support of the aforementioned hypotheses. Apart from the objective measures, we take into account the subjective measure in form of 'ease-of-estimation' rating of each of the complementary task category, as will be explained later.

4.1 Research question and hypothesis

Our study is an approach to address 3 main research questions as mentioned below:

- **Is the user perception affected by any factors while estimating correlation using scatter plots?**
- **Can the accuracy of correlation judgment be modeled using Weber's law?**
- **Is the JND dependent on the reference correlation (positive or negative) from which the difference is calculated?**

To answer these questions, we have to formulate the hypotheses for different factors affecting human perception as well as for modeling **Weber's law** and evaluating **JND** dependence on reference correlation value. We shall later evaluate them during analyses. The factors examined in our study include **density, distribution, reflective asymmetry** and **progressive symmetry**. The meaning of each of these in

context of our study will be explained in next section where we give details of each individual task and its objective.

We propose six hypotheses covering all the three research question in detail. They are summarized as follows:

- ***H1****: The hypothesized JND value is dependent on the reference correlation value.*
- ***H2****: The user accuracy of correlation judgment in scatter plot can be modeled using Weber's law.*
- ***H3****: The user perception of correlation in scatter plot is dependent on change in distribution.*
- ***H4****: The user perception of correlation in scatter plot is dependent on change in density.*
- ***H5****: The user perception of correlation in scatter plot is dependent on change in reflective asymmetry.*
- ***H6****: The user perception of correlation in scatter plot is dependent on change in progressive symmetry.*

We will test all of the aforementioned hypotheses using a controlled empirical study to derive concrete conclusions. The empirical study will be mainly divided into six broad task categories as explained in the next section.

4.2 Tasks

We design six separate tasks namely, **JND, Weber, Distribution, Density, Reflective Asymmetry** and **Progressive Symmetry** to address the three research questions mentioned previously. The first two tasks address the second and third research questions respectively while next four task categories are dedicated to resolving first research question. We explain each of of them in the following subsections.

4.2.1 Task: JND

The aim of this task is to determine whether the calculation of JND in correlation values for a sample data is dependent on the reference correlation points chosen along with generating the hypothesized average JND value. The theory related to JND (Just Noticeable Difference) is explained previously in Section 2.3. This task is designed to evaluate "precision" in user performance where "precision", refers to ability of participants to detect the correlation difference between two scatter plots, even if

they are unaware of the actual numerical correlation value. It seems naturally obvious that a difference in numerical correlations of the two scatter plots (d), which is greater than 0.35 is easily detectable by humans. Hence, we consider in our study only those differences that are less than 0.35 to infer JND.

Our experimental procedure has four reference correlation points ($0.3, 0.7, -0.3$ and -0.7), six difference values ($0.05, 0.1, 0.15, 0.25, 0.3$ and 0.35) and two *approach* conditions (above and below). With an "above" approach, the participant is given one scatter plot having one of the reference correlation r_{ref} and another with a correlation value higher than r_{ref}. Reference correlation points 0.3 and -0.7 follow the "above" approach. Conversely, with a "below" approach, the participants would be given a scatter plot with one of the r_{ref}, and another that has correlation value lower than r_{ref}. Reference correlation points -0.3 and 0.7 follow the "below" approach.

Following Harrison *et al.* (2014), we use the same adaptive psychophysical method, a "staircase procedure" to study whether inferred JND is dependent on reference correlation value, but with a slight variation. In their study, the participants were given two visualization stimuli (in our case scatter plots) and were asked to choose the one which they perceive to have a higher correlation. Instead, in our study, we ask the participants to provide numerical estimates of correlation values for each of the two scatter plots. Then instead of calculating "accuracy", we calculate "precision" by comparing their estimates of the two scatter plots to check whether the scatter plot with higher subjective correlation (participant's intuitive estimate) has indeed the higher objective correlation as well.

Unlike all the other tasks, we present the scatter plots for this task in pairs for users to make comparison and detect difference, if any, between the two. The cloud shape for all the scatter plots of this task are elliptical, as illustrated in Figure 4.1.

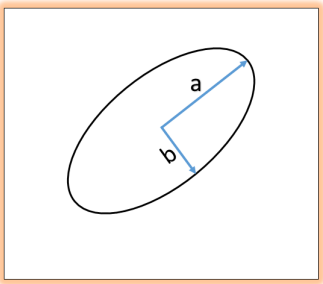


Figure 4.1: Scatter plot cloud shape for task JND, Weber and Density

To generate scatter plots having a target correlation value (r) and thus a different width of the ellipse in the cloud shape each time, we vary the value of semi-minor axis (b) of the ellipse (see Figure 4.1). Ideally if b increases, correlation will decrease and vice versa.

4.2.2 Task: Weber

The aim of this task is to determine whether the accuracy of the correlation judgment in scatter plots can be modeled using Weber's law which was explained in Section 2.4. It means using Weber's formula to predict subjective estimate based on the objective correlation value. We measure user estimates for 15 different correlation values, five in the negative direction ($-0.9, -0.7, -0.5, -0.3$ and -0.1) and ten in the positive direction $(0, 0.1, .., 0.9)$. The cloud shape for all the scatter plots of this task are elliptical, similar to that for the previous task, as illustrated in Figure 4.1.

To generate scatter plots having a target correlation value (r) and thus a different width of the cloud shape each time, we vary the value of semi-minor axis (b) of the ellipse (see Figure 4.1). Ideally if it increases, correlation will decrease and vice versa.

Finally, the difference in actual correlation value of scatter plot and user estimated value is stored which is later used for analyses as will be explained in Chapter 5.

4.2.3 Task: Distribution

The aim of this task is to observe whether the user perception of correlation is dependent on distribution of data points in the scatter plot. The point cloud shape of the scatter plot is unlike the standard one and the objective is to identify whether participant's estimates are biased towards them. We examine their effect by distributing points along the 45 degree and -45 degree line, such as the one shown in Figure 4.2.

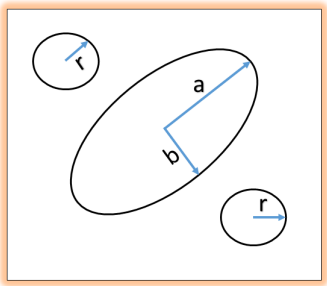


Figure 4.2: Scatter plot cloud shape for task Distribution

We distribute all the data points equally along the -45 degree line in two circles of equal radius and along the 45 degree line in an ellipse. The variation in distribution is introduced by altering the number of data points along each line while at all times ensuring the two circles have equal number of points. Ideally, as the number of points along the 45 degree line increases, the correlation value will increase and vice versa. For instance, we can consider the circles as outliers which diminish the effect of positive correlation rendered by the ellipse. As more points are distributed in the circles, the distribution on the -45 degree line increases and the effect of the outliers over shadows the positive correlation effect formed by ellipse. Thus, we obtain a

less positively correlated scatter plot with each distribution level, which eventually becomes negative when maximum data points are distributed in the two circles. The design of this task and the various distribution ratios considered are explained further in Chapter 5.

4.2.4 Task: Density

The aim of this task is to observe whether human perception is affected by the number of points used to plot a scatter plot, henceforth denoted by the term "density" for the purpose of this study. In our study, we focus on five density levels - $40, 60, 80, 100, 120$- which correspond to very small, small, medium, large and very large density level, respectively. We strategically choose three correlation values - 0.7, 0.5, 0.3- to analyze density variations at high, neutral and low correlation values, respectively. Each of these three would be plotted with the five different densities mentioned above. For this task, the scatter plot has an elliptical cloud shape, similar to the one used in Task - Weber as illustrated in Figure 4.1.

Ideally, it is possible to increase or decrease the number of data points while simultaneously keeping the correlation constant. For example, the Pearson correlation coefficient for both the dataset in Table 4.1 are the same and equal to 1. We can always scale the given data points in either directions to generate additional or fewer data points as done in the stated example. Having said that, we refrain from using very large sample sizes (> 120) as they contribute to visual clutter and the data points tend to overlap with each other due to which the visual density appears lower than what it really is. Similarly we avoid plotting scatter plots with scarce data set (< 40) as the cloud shape becomes insignificant and unidentifiable.

Data point	X	Y
1	10	10
2	20	20
3	30	30
4	40	40
5	50	50

((a)) 5 data points with Pearson correlation coefficient (R) = 1

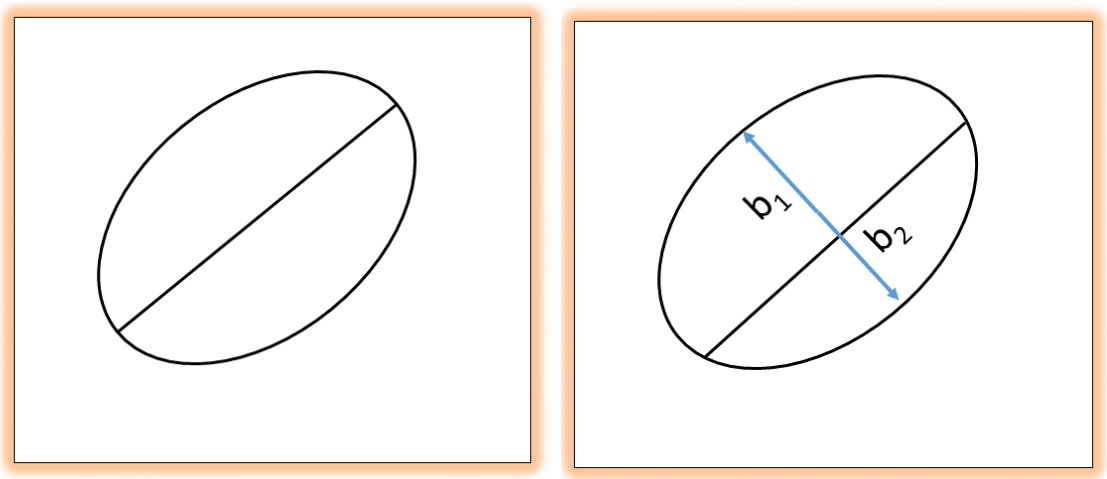
Data point	X	Y
1	10	10
2	20	20
3	30	30
4	40	40
5	50	50
6	60	60
7	70	70
8	80	80
9	90	90
10	100	100

((b)) 10 data points with Pearson correlation coefficient (R) = 1

Table 4.1: Different density data sets with common Pearson coefficient of correlation

4.2.5 Task: Reflective Asymmetry

The aim of this task is to observe the impact of variation in "reflective asymmetry" on the user's perception of correlation in scatter plot. In context of our study, "reflective asymmetry" refers to the unequal symmetry of data points on either side of the 45 degree line. A scatter plot displays no reflective asymmetry if its cloud shape is laid out evenly on either side of the line i.e. the half on one side of the 45 degree line is identical to the other half on the opposite side of the line (See Figure 4.3 (a)). On the other hand, a scatter plot cloud with reflective asymmetry, is skewed along the 45 degree line i.e. one side is bigger than the other on the opposite side of the line (see Figure 4.3 (b)).



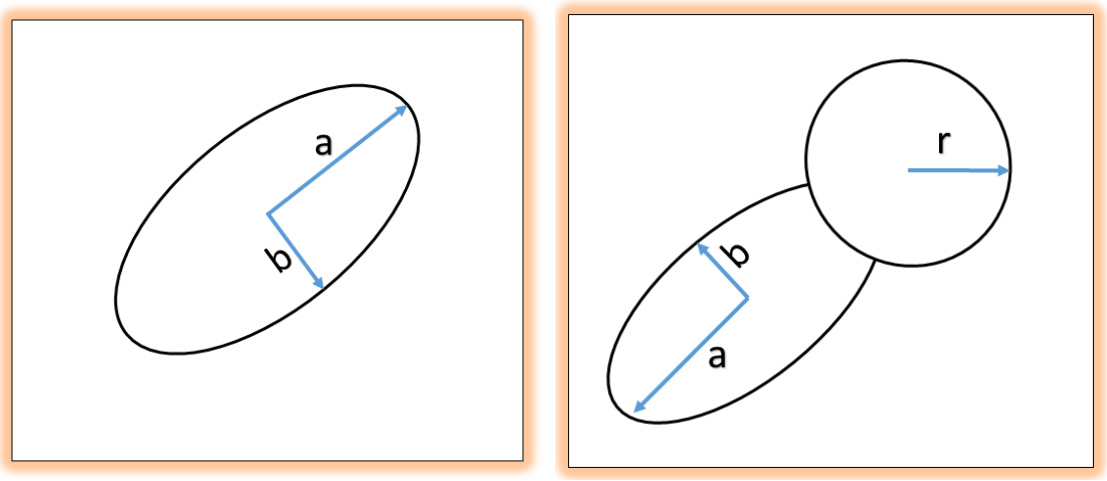
(a) No Refractive Symmetry (b) Absolute Refractive Symmetry

Figure 4.3: Scatter plot cloud shape for task Reflective Asymmetry

In this task, the scatter plots have cloud shape as either of the two graphs shown in Figure 4.3. It is composed of two half ellipses, one on either side of the 45 degree line. The variation in refractive symmetry is introduced by varying the values of the semi minor axes of the two half ellipses such that the correlation remains constant. We begin with equal size of both b_1 and b_2 and gradually increase b_1 and decrease b_2. For this task, we consider three correlation values $0.5, 0.7$ and 0.9. Each has three cases, one where there is no reflective asymmetry and other two for different reflective asymmetry levels. The case where there is no refractive symmetry observed, values of both the semi-minor axes become equal. These are used as a benchmark to analyze the transition in user perception from introduction of reflective asymmetry in other two cases.

4.2.6 Task: Progressive Symmetry

The aim of this task is to observe the impact of variation in "progressive symmetry" on the perception of correlation in scatter plot. In context of our study "progressive symmetry" refers to the concentration of data points in two clusters (an ellipse and a circle at its top) along the 45 degree line (see Figure 4.4 (b)). A scatter plot has no progressive symmetry if the scatter cloud is formed entirely from data points inside the ellipse only (see Figure 4.4 (a)).



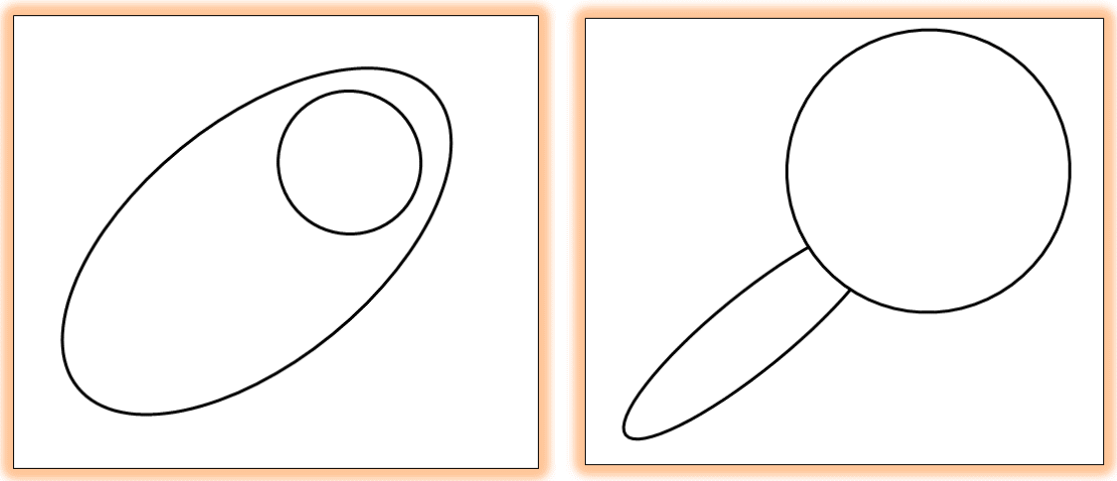
(a) No Progressive Symmetry (b) Absolute Progressive Symmetry

Figure 4.4: Scatter plot cloud shape for task Progressive Symmetry

In this task, the scatter plots have cloud shape as either of the two graphs shown in Figure 4.4. It is composed of an ellipse, along the 45 degree line and a circle overlapping the ellipse towards the end of line. The variation in progressive symmetry is introduced by varying the values of the semi minor axes of the ellipse (b) and radius of the circle (r) such that the correlation remains constant at 0.7. We consider six different variations of progressive symmetry, one where there is no progressive symmetry and other five for varying progressive symmetry levels. In case of absence of progressive symmetry, the cloud shape is formed entirely of the ellipse only. These are used as a benchmark to analyze the transition in user perception from introduction of progressive symmetry in other five cases.

In order to keep the correlation value constant, we must increase the radius of the circle and decrease the semi minor axis of the ellipse (or vice versa) simultaneously. For instance, in absence of the circle ($r = 0$), all points are distributed evenly inside the ellipse (see Figure 4.4(a)). If a small circle which completely fits inside the ellipse is introduced, the points towards the end of the ellipse tend to squeeze and come closer to each other (see Figure 4.5(a)) and hence the correlation value will increase slightly compared to its previous value. Now, to prevent this change in correlation, we must counterbalance the increase by introducing some negative correlation. As a

result, we increase the semi minor axis of the ellipse such that the points are farther apart from each other and thus the increase in correlation from circle is cancelled by the decrease from the ellipse and the correlation remains constant. Conversely, if the circle introduced partially overlaps the ellipse, the data points towards the end of the ellipse tend to scatter (see Figure 4.5(b)) and the correlation value decreases. Subsequently, we decrease the semi minor axis of ellipse to bring data points in close proximity of each other and thus balance out the decrease brought by circle such that correlation is maintained constant.



(a) Contraction of data points in circle which brings data points closer and thus increases correlation

(b) Expansion of data points in circle which spreads the data point farther and thus decreases correlation

Figure 4.5: Variations of Progressive Symmetry keeping correlation constant, $R = 0.7$

4.3 Variables in Experiment

Similar to most empirical studies, several variables may potentially effect the user performance. Hence we focus our study with a small number of variables while controlling the rest of them. A **dependent variable** is measured each time the **independent variable** is changed while simultaneously keeping the **control variables** constant during the experiment. This is done to ensure least possible confounding effect and generation of comparable results. The following subsections will explain independent, dependent and control variables in our study.

4.3.1 Task Variables

Task variables, henceforth, will be used to refer to all the independent variables that will be used specifically for each of the six task categories. We have 5 main task

variables which will take different values depending on the task category, as explained below:

Data Points (denoted by DP): It refers to the number of points plotted on the scatter plot. It will take on values such as $40, 60, 80, 100$ and 120.

Distribution Ratio (denoted by DR): It refers to the distribution of data points in a scatter plot between a circle, ellipse and another circle of same radius i.e. circle:ellipse:circle. It will take on values such as 5:70:5, 10:60:10, etc.

Semi major axis of Ellipse (denoted by a): It refers to the length (in units) of the semi major axis of an ellipse in the scatter plot.

Semi minor axis of Ellipse (denoted by b): It refers to the length (in units) of the semi minor axis of the ellipse in the scatter plot.

Radius of Circle (denoted by r): It refers to the radius (in units) of the circle in the scatter plot.

4.3.2 Control Variables

There are a few universal control variables in this study, which are kept constant throughout the experiment, while some others are specific to a particular task category. Since these can strongly influence other values, we force them to remain unchanged to study the relative impact of independent variables. Below we explain the universal control variables and task specific control variables in brief.

4.3.2.1 Universal Control Variables

We take into account three main control variables which remain fixed throughout the study to ensure consistency.

Dimension of SCPs: All the scatter plot used throughout the study are similar in format, design and dimensions. The details of their design will be explained in Chapter 5. All these are kept constant so that participant can focus entirely on the cloud shape of the scatter plot in question.

Layout of SCPs and slider bars: We strategically place the scatter plots and their corresponding slider bars (used to record user estimate of correlation value in the scatter plot) on the screen at the same position every time to reduce any confounding effect due to change in location of scatter plot. This is because the participant may need to move their eyes in different directions every time which might be distracting.

Position of the 'Next' button: The 'Next' button, used to transport the participant from one trial to the next is placed at a constant position on the screen. This is done to reduce confounding effect arising since the participant may need to move the mouse in different amounts to different positions.

	Task	Independent Variable	Task Specific Control Variables
1	Task Weber	b	$DP = 80$ $DR = 0 : 80 : 0$ $a = 7$ $r = 0$
2	Task JND	Reference Correlation Difference in stimuli b	$DP = 80$ $DR = 0 : 80 : 0$ $a = 7$ $r = 0$
3	Task Distribution	DR	$DP = 80$ $a = 7$ $b_1 = 7$ $r = 0$
4	Task Density	DP Reference Correlation b	$DR = 0 : 80 : 0$ $a = 7$ $r = 0$
5	Task Reflective Asymmetry	b_1 b_2 Reference Correlation	$DP = 80$ $a = 7$ $DR = 0 : 80 : 0$ $r = 0$
6	Task Progressive Symmetry	b r	$DP = 80$ $a = 7$ $DR = 40 : 40 : 0$

Table 4.2: Experiment variables according to their task category. Here, DP refers to the density, DR is the distribution ratio, r is the radius of the circle and b is the semi-minor axis of the ellipse in the scatter plot (b_1 and b_2 are semi-minor axes of two half ellipsoid for task Reflective Asymmetry

4.3.2.2 Task Specific control variables

While taking care of universal control variables ensure consistency throughout the experiment, we will introduce some task specific control variables to maintain consistency within each task too. These are selected from within the set of Task variables described previously and kept constant. Table 4.2 gives an overview of these.

4.3.3 Independent Variables

The aim of our study is to observe the several factors which may potentially affect the perception of correlation using scatter plots. Hence, we have independent variables specific to each of the six task categories as explained in Table 4.2.

The different variants of each of the independent task specific to each of the task category will be explained under User Study Design in Chapter 5.

4.3.4 Dependent Variables

Accuracy is one of the most common dependent variables used in empirical studies. In our study, we use it to measure the impact of different factors affecting perception of correlation in scatter plots for different task categories. Accuracy indirectly signifies the intensity of the impact, the independent variables have on the perception of correlation.

It will be measured as the *signed* (positive or negative) difference between the estimated and the actual correlation value of the scatter plot. We will use the absolute error value when performing statistical analyses whereas signed error value will be useful for determining whether and by how much on an average, the participant overestimate or underestimate the actual correlation value. A negative sign is symbolic of underestimation whereas a positive sign denotes overestimation. For example, if estimated $R = 0.3$ and actual $R = 0.5$, the accuracy would be stores as: $0.3 - 0.5 = -0.2$. The negative sign of -0.2 indicates that the participant estimated a value, lower than the actual value. Lastly, an accuracy of 0 indicates the participant estimated correctly and the actual and estimated correlation value are a perfect match.

For correlation estimation task, we have an option to follow input procedure of previous studies and allow user to type in the answer but it can create typographical error. Also, we refrain from using radio buttons to provide optional answers for users to select their estimate of correlation as it results in *leading effect* which is a confounder in that it '*leads*' the participant to the correct answer by confining his range of answer choices . Thus, among these two extreme options, we devised a third option of collecting user response using a slider bar. The participants can slide it left and right to record their answer. Also, the amount of time taken to adjust slider bar is insignificant and can be infinite since we do not statistically analyze response time to draw any inference.

4.4 Measurement Metrics

Our study involves two measurement metrics: objective measure and subjective measure.

4.4.1 Objective Measure

To analyze user performance, we include **accuracy** as the objective measure in this study. It will be used to analyze the effect of different factors such as density distribution on user perception of correlation using scatter plots.

4.4.2 Subjective Measure

In this study we will use the the **familiarity** and **ease-of-estimation** rating as the subjective measure for user performance. Before the start of the study, we acquire the participant's level of familiarity with scatter plots. Once the study has ended, we provide them with a feedback form which will compare two different categories of scatter plots at a time. The participant must specify for which of the category it is relatively easier to estimate correlation.

For both the rating techniques, we will use the five-level Likert scale, which was introduced by Likert. It offers a fixed choice of responses to measure the level of agreement or disagreement with a particular statement or question and provides a method of ascribing quantitative value to qualitative data, thus making it amenable to statistical analysis.

In our study, for familiarity rating, the Likert scale provides a choice of 5 responses which include *Not at all familiar, Slightly familiar, Moderately familiar, Very familiar* and *Extremely familiar*. In case of ease-of-estimation rating, the choices are *Category 1 : much easier, Category 1: easier, Same, Category 2: easier* and *Category 2: much easier*. A numerical value is assigned to each potential choice which we will average towards the end to estimate the overall level of familiarity and ease of estimation of correlation in scatter plots.

4.5 Techniques for Analyses

To analyze the six hypotheses provided in Section 3.2, we will analyze the accuracy and precision in user performance for the six tasks separately. Within each task, we also analyze the user performance based on variation in value of controlled variable, as explained below.

4.5.1 Task Analyses

Below we present the details of analyses of each of the six task categories and the controlled variable used for each analyses. We provide illustrations for each task to show hierarchy of the analyses process.

Task: JND

Figure 4.6 illustrates the divisions of the performance analyses in task JND. We will look for patterns in user performance within the positive reference correlation points i.e. R=0.7 and R=0.3. Similar procedure will be followed for negative reference correlation points i.e. R= -0.7 and R= -0.3. Later, we will collectively analyze all four to conclude if one common JND value can be deduced for our dataset.

Task: Weber

Figure 4.7 illustrates the divisions of the performance analyses in task Weber. For this task, we will analyze and compare the accuracy in user estimation of positive correlation to figure out the existence of some range over which the accuracy is the high. We then compare positive correlation in relation to its negative counterpart. We will examine any pattern observed and derive conclusions based on that.

Task: Distribution

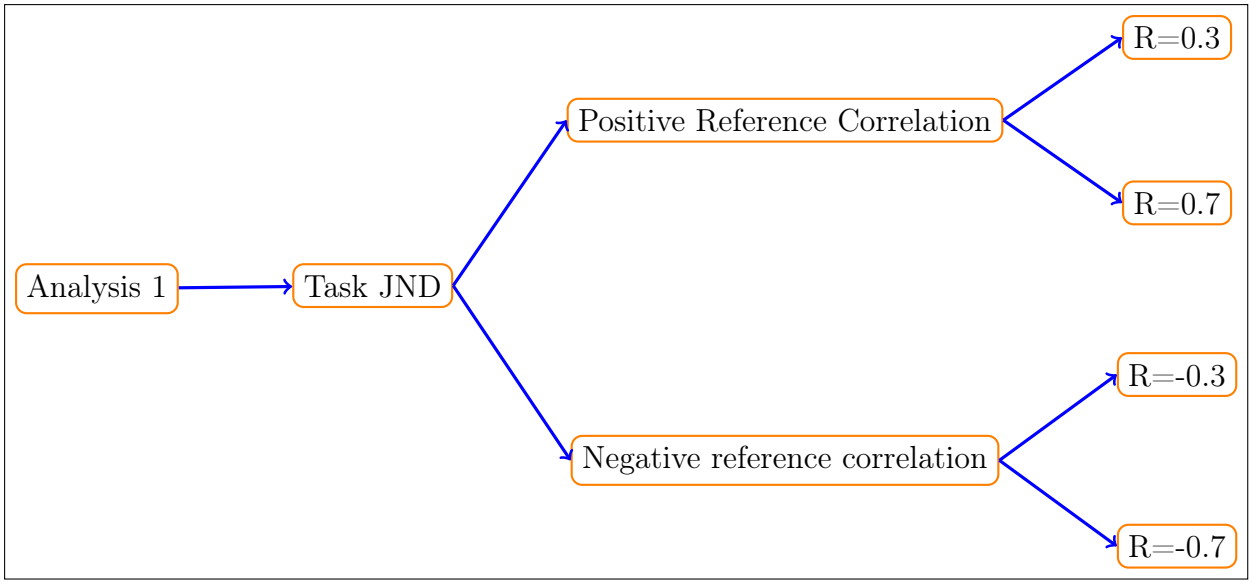


Figure 4.6: Hierarchy of result analysis for task JND

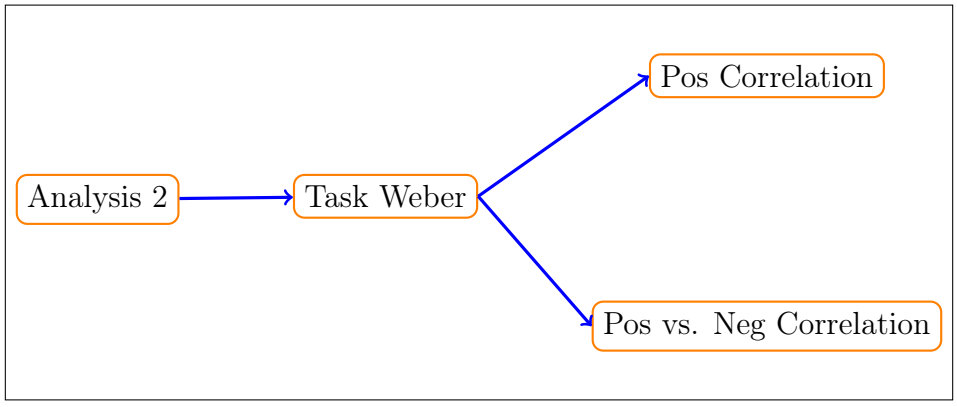


Figure 4.7: Hierarchy of result analysis for task Weber

Figure 4.8 illustrates the divisions of the performance analyses in task Distribution. We will analyze the effect of variation in distribution ratio on user performance.

Task: Density

Figure 4.9 illustrates the divisions of the performance analyses in task Density.

We will analyze the user performance for each of the 5 density values - 40, 60, 80, 100 and 120. In addition to this, we also aim to analyze performance within each density group to determine the correlation value (high=0.7, low=0.3 or neutral=0.5) for which we obtain the most accurate user performance.

Task: Reflective Asymmetry

Figure 4.10 illustrates the divisions of the performance analyses in task Reflective Asymmetry. We will analyze the effect of varying the reflective asymmetry of the cloud shape in the scatter plots with respect to the three reference correlation values (0.3, 0.5 and 0.7). We also compare within each reference correlation, the user performance with respect to level of reflective asymmetry levels ($=, >, \gg$).

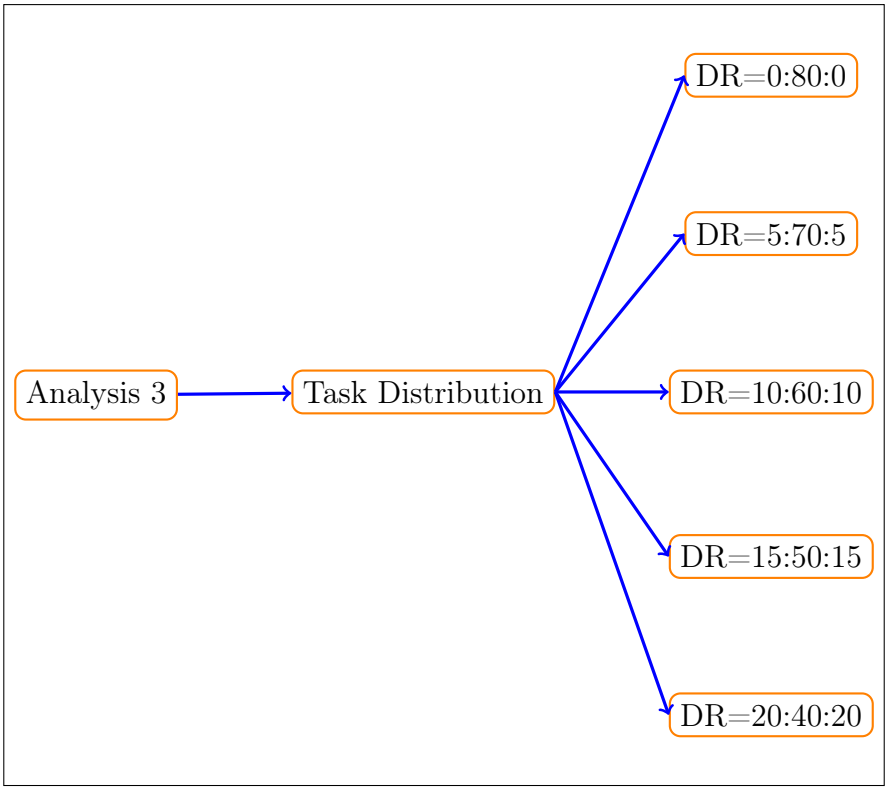


Figure 4.8: Hierarchy of result analysis for task Distribution

task: Progressive Symmetry

Figure 4.11 illustrates the divisions of the performance analyses in task Progressive Symmetry. We analyze whether there is an effect on user performance if there is a change in progressive symmetry by varying the semi-minor axis and radius of the circle with respect to each other.

4.5.2 Non-Parametric Test in SPSS

As explained in Section 2.5.4.1, we use **Friedman analysis** as the preliminary analysis test to detect presence of overall significance in the data set. It will be followed by pairwise comparison of groups using **Wilcoxon Signed-Rank test** to detect the source of this significance difference. The significance level of these two analyses will be set to $\alpha = 0.05$, as discussed in Section 2.5.4. In this study, we will use SPSS, one of the predictive analytics software, to perform both analyses.

In our study, we report all the results, regardless of the presence of a significance or not. If there is a significant difference in the variations of independent variable in a task ($p < 0.05$), the accuracy in user performance in estimation of correlation will be ranked in an increasing order of mean rank with a higher rank denoting less error rate and lower rank denoting higher error rate, except in case of task JND whether we do not analyze the error but the ability to detect a stimuli difference and hence

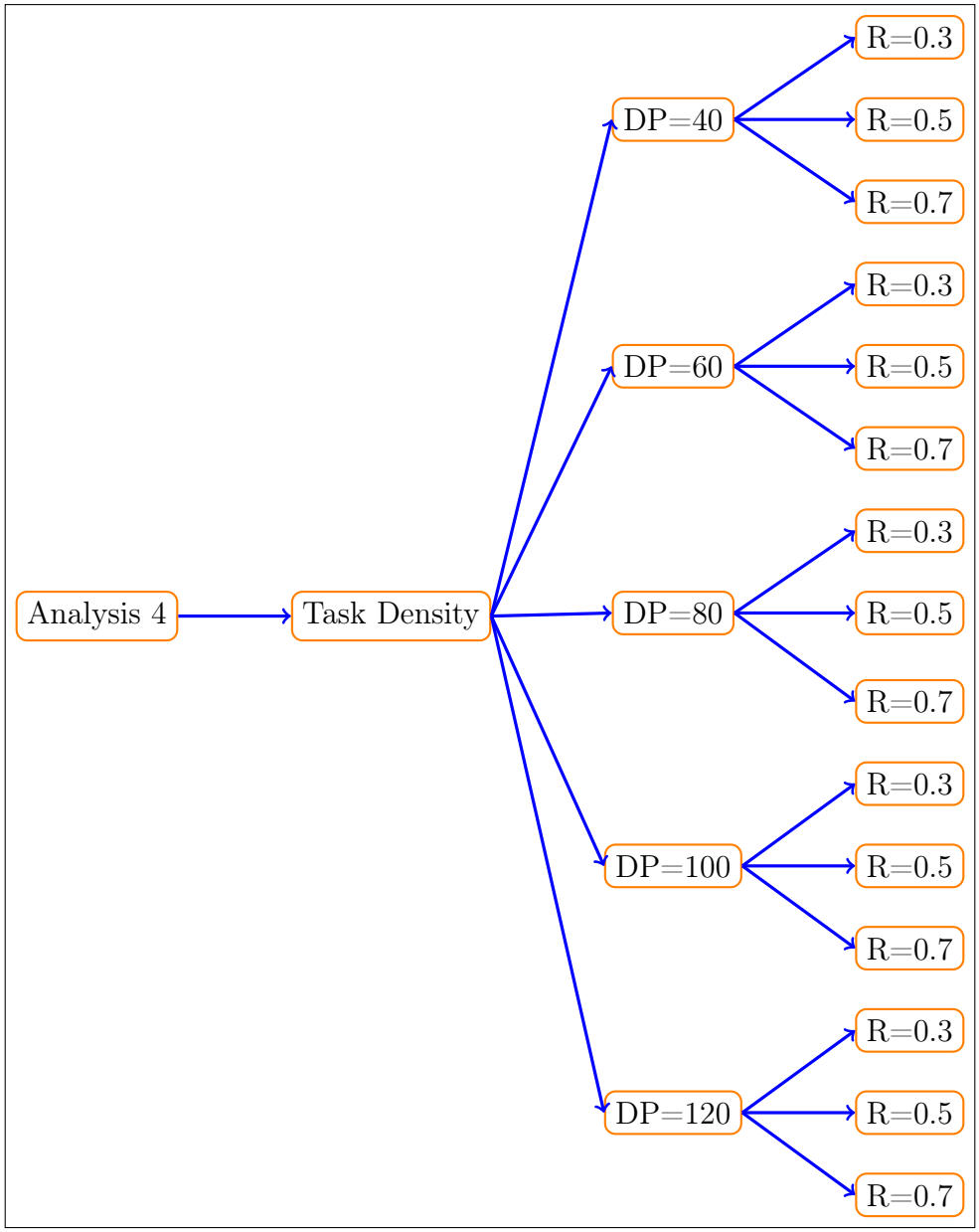


Figure 4.9: Hierarchy of result analysis for task Density

there a high rank means better user performance.

We explain in the next chapter, the design of the various task categories and the actual implementation will be explained in subsequent chapters.

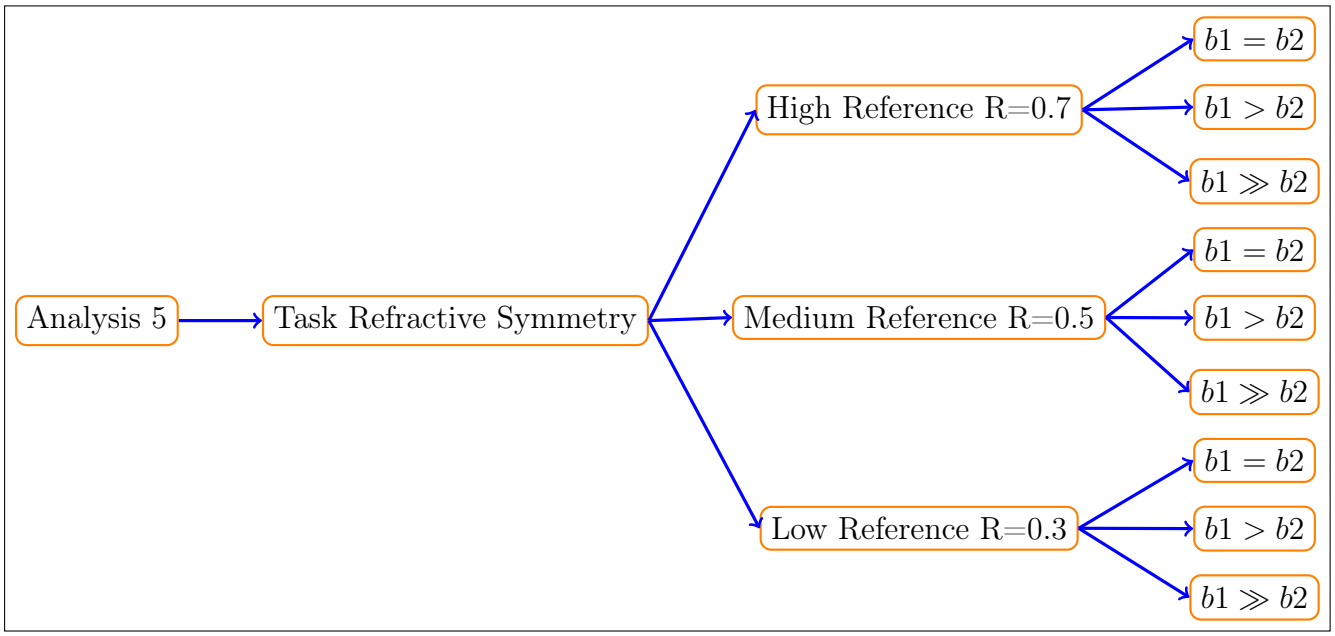


Figure 4.10: Hierarchy of result analysis for task Reflective Asymmetry

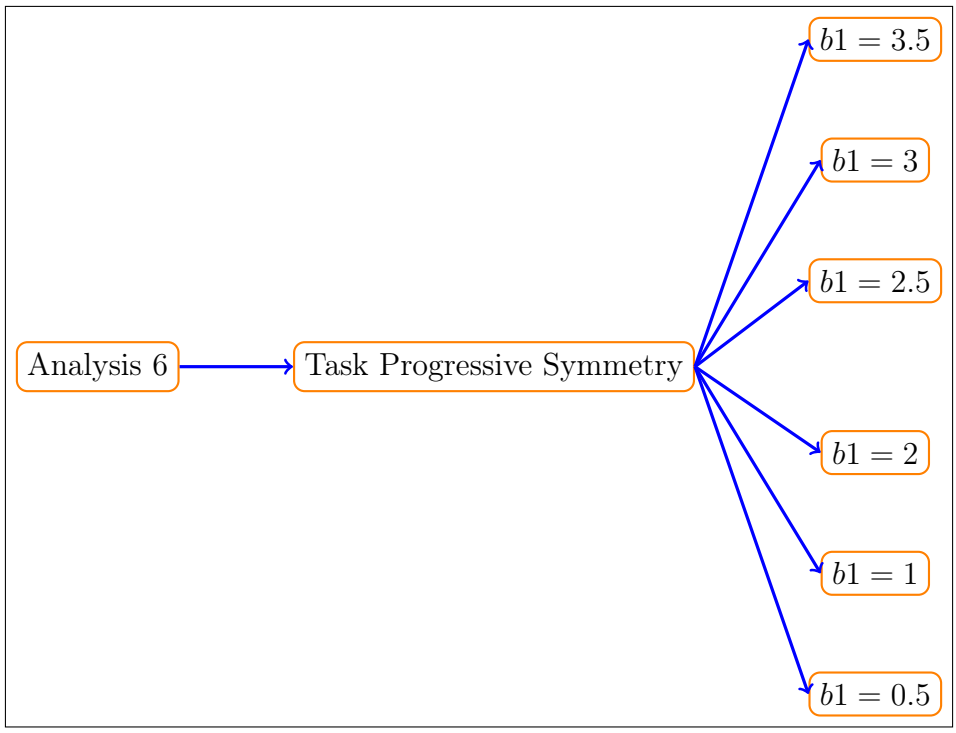


Figure 4.11: Hierarchy of result analysis for task Progressive Symmetry

Chapter 5

User Study Design

User design study is the process of systematic development of an experiment such that we gather all the relevant experiment data (user results) to the best of our ability, reducing any biases involved & taking measures to reduce the confounding effect, if any. These confounding effects are not related to the study, per se, but might affect the outcomes. Hence it is a crucial part of an empirical study, which might be hampered if we fail to carefully design it taking into account the proposed hypothesis, required metrics, systematically designed procedure & resultant conclusions.

In this chapter, we focus on 6 main design elements - task design, stimulus design, software design, pre-study training design, pre-study presentation design and feedback form design. We begin with a brief overview before delving into the details for all these in subsequent subsections.

5.1 Overview

In the following subsections, we briefly describe the different task categories, the trials associated with each and the user interaction with the software.

5.1.1 Tasks

We have a total of 6 task categories for the user study namely,task JND, task Weber, task Distribution, task Density, task Reflective Asymmetry and task Progressive Symmetry. Each of these is aimed at observing the various factors that affect the perception of correlation using scatter plots. Section 5.3 explains each task and the relevant stimulus design in detail.

5.1.2 Trials

Trials refers to the screen presented to the participants during the study. All the trials, will contain two scatter plots as illustrated in Figure 5.1 and measures user's performance while they estimate correlation.The trials are not categorized or paired by the task category that they belong to, as it might result in *leading effect*. Hence, throughout the course of the study, the participants remain oblivious to the task category a particular Scatter Plot is tested for. The randomization of trials is further explained in Section 5.4.3.

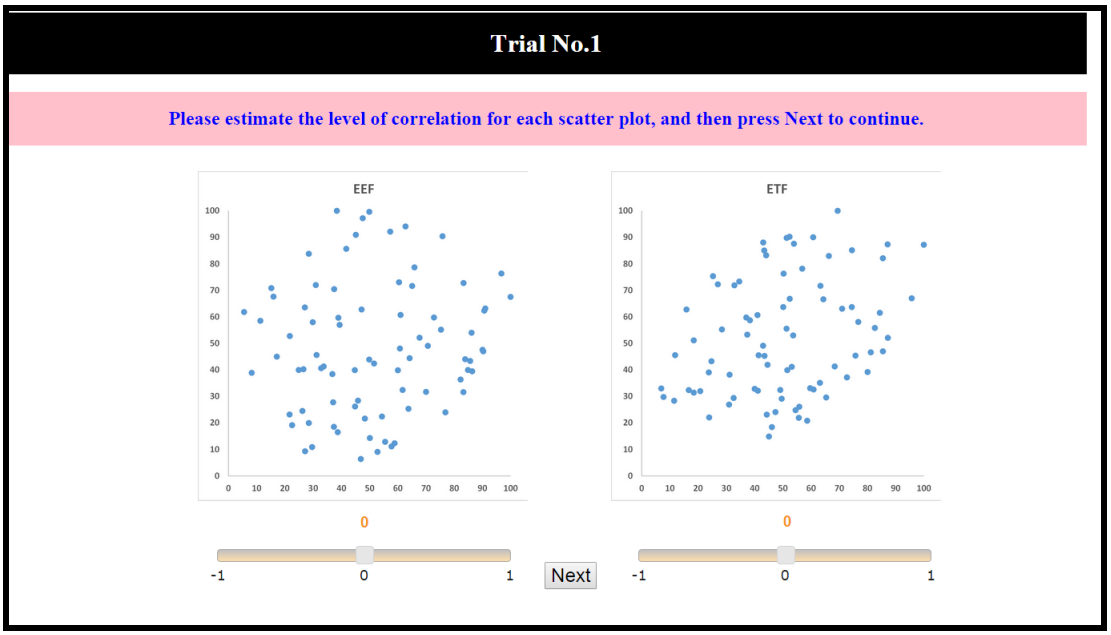


Figure 5.1: Trial screen in software

5.1.3 Software Interaction

The participants will be performing the user study on computer system that have the software pre-installed on them. The software will present them with trials and participants must record their estimate of correlation for each scatter plot. To collect user's estimate of correlation, we provide a horizontal slider bar which has a single handle that can be moved in either directions using a mouse (see Figure 5.2).

It is initially set to default value 0 in the middle of slider bar. It has the same scale range as that of mathematical formula calculated correlation value i.e. -1 to $+1$ with tick marks at -1, $+1$ and 0 values and incremental step size of 0.05. The slider bar can be shifted left or right to make a negative or positive correlation estimate respectively. After much introspection, this method of collecting user response seems most efficient as its less error prone compared to its counterparts - text boxes, radio buttons, etc. Also, it is self-explanatory, requires little learning effort and is familiar

to most of the participants. The participants can also make use of the keyboard to interact with the software, as shown in Figure 5.2.

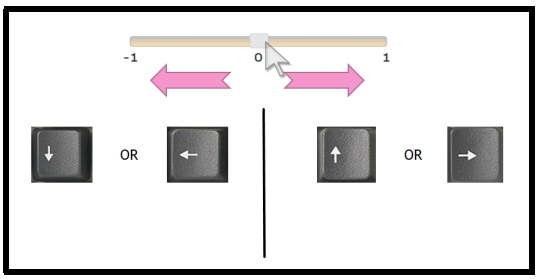


Figure 5.2: User interaction with software

Stimuli generation and organization is described in the next section while Section 5.3 is dedicated to the design of stimulus for each task. We then discuss the software implemented to collect user data in section 5.4 and conclude this chapter with design of pre-study presentation and feedback survey form in section 5.5 and 5.6 respectively.

5.2 Stimuli Design

The data points for each SCP belonging to different task categories are generated using programs written in C++ and then plotted on a graph using MS Excel's XY Scatter graphs. The generation of the data set is subjected to some rules given in Section 5.2.2. All the scatter plots have similar physical features regardless of their category but a unique label ID, which is carefully devised to conceal it from the participants, details of which will be given in Section 5.2.3 . Some of the stimuli, henceforth called *checkpoints* are for sole purpose of validating the user data while some of the stimuli overlap in the sense that the same stimuli could be used for more than 1 task's graph calculations. The details of stimuli organization and re-utilization will be provided in section 5.2.1.1..

5.2.1 Stimuli organization

We explain in each of the following section, the actual number of stimuli needed for each task, their presentation during the main trial and their repetition during course of the experiment.

5.2.1.1 Main Trials

Below we present a categorization of the stimuli used for the study. We have 80 stimuli for task JND, 6 for task Weber, 6 for task Distribution, 12 stimuli for task Density, 6 stimuli for task Reflective Asymmetry, and 5 for task Progressive Symmetry. Task

JND is further categorized into 2 sub tasks namely- Task JND Coarse and Task JND Fine. Task JND Fine in turn has 4 sub task categories, namely, Task JND Fine 0.7, Task JND Fine 0.3, Task JND Fine -0.7 and Task JND Fine -0.3. Each of these tasks have 24, 14, 14, 14 and 14 stimuli respectively which gives a total of 80 Task JND stimuli. The study is carefully planned to ensure an adequate number of stimuli for each task category, some of which are repeatedly measured in accordance with **principal of repeated measure**.

Repeated Measures are used when the research plan exposes the same set of participants to several measurement occasions in different times without any manipulation (Tenenbaum and Driscoll, 2005). Thus the simplest repeated measure design is when 1 or more measures are taken for the same subject in a *counter-balanced* order. The 2 major reasons for us to use repeated measures are:

- increase in statistical power and hence fewer participants are required.
- eliminate errors (faulty handling of the slider bar, misjudgment of correlation, etc) made by user in the first instance.

As part of our study, we repeat each stimuli from all task categories (except Task JND) *at least* 3 times whilst ensuring they have adequate gap between them when displayed to participants in actual trials. Table 5.1 lists all the stimuli used along with the multiple task category they are used for.

Hence, we now get the total number of stimuli for each task category as :

- Task JND : 24 + 14 + 14 + 14 + 14 = 80
- Task Weber : 6 * 3 = 18
- Task Distribution : 6 * 3 = 18
- Task Density : 12 * 3 = 36
- Task Reflective Asymmetry : 6 * 3 = 18
- Task Progressive Symmetry : 5 * 3 = 15

Adding all of them gives a total of 185 stimuli. Furthermore, we have 10 additional stimuli for training purpose which are presented to user before start of the Testing session as will be explained in the next section.

5.2.1.2 Training trials

We give a cumulative training session for all task categories, at the start of the experiment, rather than giving training sessions, specific to a task category. This is done

	Stimuli label	Correlation (R)	task JND	task Reflective Asymmetry	task Distribution	task Progressive Symmetry	task Density	task Weber
1	ENB	-0.9						✓
2	CECNB	-0.9			✓			
3	EAB	-0.8						
4	ESB	-0.7	✓					✓
5	CECSB	-0.7			✓			
6	ECDB	-0.65	✓					
7	ECB	-0.6	✓					
8	EPDB	-0.55	✓					
9	EPB	-0.5						✓
10	CECPB	-0.5			✓			
11	EQDB	-0.45	✓					
12	EQB	-0.4	✓					
13	ETDB	-0.35	✓					
14	ETB	-0.3	✓					✓
15	CECTB	-0.3			✓			
16	EEB	-0.1						✓
17	CECEB	-0.1			✓			
18	EEZZEE	0						✓
19	EZDF	0.05	✓					
20	EEF	0.1	✓					✓
21	EEDF	0.15	✓					
22	EDF	0.2	✓					✓
23	CECDDF	0.25			✓			
24	ETF	0.3	✓				✓	✓
25	ETFLL	0.3					✓	
26	ETFL	0.3					✓	
27	ETFS	0.3					✓	
28	ETFSS	0.3					✓	
29	ETDF	0.35	✓					
30	EQF	0.4	✓					✓
31	EQDF	0.45	✓					
32	EPF	0.5	✓	✓	✓		✓	✓
33	EPFSS	0.5					✓	
34	EPFSS	0.5					✓	
35	EPFL	0.5					✓	
36	EPFLL	0.5					✓	
37	CECPF	0.5			✓			
38	EEPFP	0.5		✓				
39	EEPFT	0.5		✓				
40	EEPFS	0.5		✓				
41	EPDF	0.55	✓					
42	ECF	0.6	✓					✓
43	ECDF	0.65	✓					
44	ESF	0.7	✓	✓		✓	✓	✓
45	ESFL	0.7					✓	
46	ESFLL	0.7					✓	
47	ESFS	0.7					✓	
48	ESFSS	0.7					✓	
49	EESFQ	0.7		✓				
50	EESFDD	0.7		✓				
51	EESFP	0.7		✓				
52	ECSFQ	0.7				✓		
53	ECSFT	0.7				✓		
54	ECSFE	0.7				✓		
55	ECSFD	0.7				✓		
56	ECSFP	0.7				✓		
57	EAF	0.8	✓					✓
58	EADF	0.85	✓					
59	ENF	0.9	✓	✓				✓
60	EENFDD	0.9		✓				
61	EENFD	0.9		✓				
62	EENFT	0.9		✓				
63	ENDF	0.95	✓					
64	ETBETB	0.3	✓					
65	ESFESF	0.7	✓					

Table 5.1: Stimuli table conforming to the task categories they are utilized for

so that the participant performance is independent of the task category, because in all the trials the participants have to estimate correlation for the two scatter plots.

The participants are subjected to 5 Training trials before the start of the Testing trials. Each training trial consists of two stimuli each and thus, the total 10 stimuli provide a *'learn and improve'* experience for participants. The stimuli and the order in which they are presented to the participants during the Training Session are carefully designed starting from the easily identifiable scatter plots to more difficult ones as illustrated in Figure 5.3.

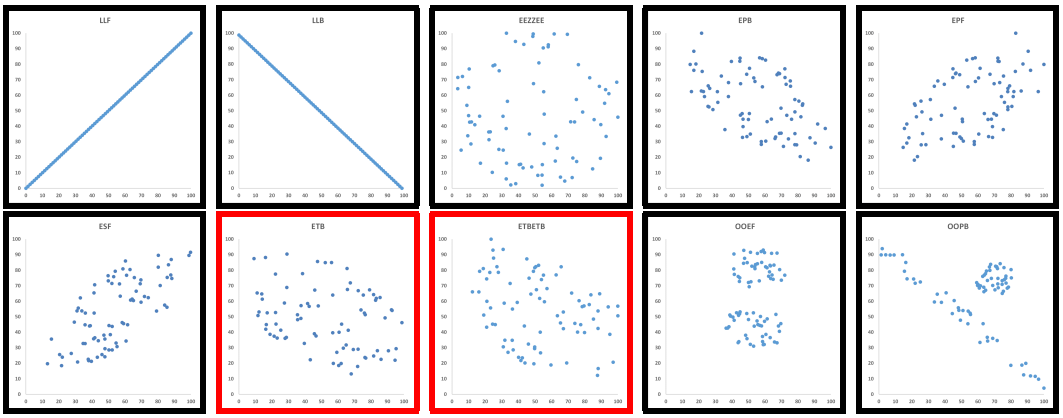


Figure 5.3: Stimuli for Training session ordered by increasing level of difficulty from left to right. The highlighted scatter plots represent equal correlation but different point cloud shape

Due to the learning curve of some of the participants and the subsequent confounding effect, we refrain from re-using the stimuli presented during the training session as part of Testing sessions. Instead we generate new data set for testing session stimuli.

Training session solve the dual purpose of familiarizing the participants with the study alongside checking whether their data is usable. The latter can be achieved by examining the training data of each participant before including it in statistic calculation. The data is deemed fit for use only if:

- for at least 20% (first two) of the stimuli, the estimated correlation values match the actual ones, **and**
- for at least 60% (next six) of the stimuli, the absolute different between estimated and actual correlation is less than or equal to 0.3.

Failure to achieve these set standards reflect that the participant lacks the basic understanding of scatter plots and correlation, in general, and thus their responses would be mostly random guesses. Along with scrutinizing the training data, we need to ensure the participant performs efficiently throughout the experiment, and thus we introduce *check points* as explained in the next section.

Organization of Training stimuli

We begin by presenting the two *easiest* SCPs having correlation +1 and -1 during the first Testing trial. This helps to boost the participants confidence level. We expect all the participants to estimate it correctly as these have been explained in the pre-study presentation too.

We next present participants with comparatively less easier stimuli having correlation 0 & -0.5 and 0.5 & 0.7 in second and third training trial respectively.

We purposely design the fourth trial to displays two SCPs of exactly same correlation value but different data set and hence a slightly different cloud shape (see scatter plots highlighted in red in Figure 5.3). This is to provide an example of **many-to-one mapping** property. It also helps break the cliche that two SCPs in a trial cannot have same correlation value. This proves beneficial in Testing trials having two SCP of exactly same mathematical correlation value but significantly different cloud shape, adjacent to each other.

Lastly, we display two stimuli having non-uniform cloud shape, but different to the one actually used in the testing trials. This allows users to adapt to the possibility of appearance of unusual SCP that have cloud shapes different from the standard ones (straight lines and ellipses) during the main trials. This is necessary because stimuli for task Reflective Asymmetry, Progressive Symmetry and Distribution make use of such cloud shapes.

For each Training trial, the participant has provision for 'Check my answer' button which pops up an alert box on top of the screen, for them to compare their answers to the actual formula-calculated correlation coefficient for both the stimuli. Figure 5.4 illustrates an example of Training session trial giving feedback to participant's estimates of correlation.

5.2.1.3 Re-using stimuli for statistical calculation

Some of the stimuli are unique to a particular task category whereas others aren't. We re-utilize some scatter plots from task JND for statistical calculations of several other tasks which magnifies the data available for analyses.

Instead of generating new scatter plots, some of the stimuli used in task JND for correlation, $R = -0.7, -0.3, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9$ are utilized again for statistical calculation of task Weber as well since both tasks have the same point cloud shape and same measurement of metrics used to plot the scatter plot for corresponding correlation. While 'precision' is calculated using value of R in task JND, the same values are used to calculate accuracy i.e. the difference between estimated and actual correlation values.

Similarly, for task Density with density level equal to 80 and correlation, $R =$

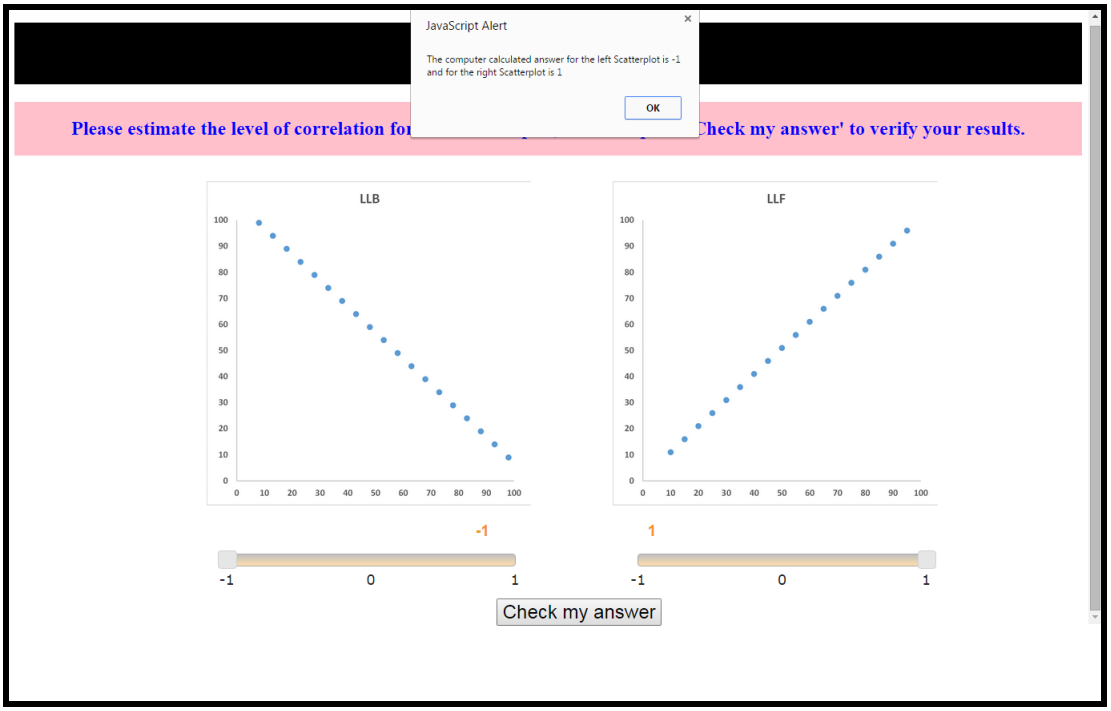


Figure 5.4: Training session screen providing feedback to the participants in the alert box at top

$0.3, 0.5, 0.7$, the user estimates from task JND are again utilized. However, since other density levels require a different density level such as 40 or 100, they cannot be re-utilized from any of the existing task categories and hence generated exclusively.

Lastly, for each of the tasks Reflective Asymmetry, Progressive Symmetry and Distribution, an independent variable level acts as a reference point, generally the one that is devoid of effects of distribution, progressive symmetry and reflective asymmetry, to compare against user performance at different levels. This helps in estimating whether there is any change in accuracy due to variation in perception abilities of participants due to variation in scatter plot point cloud. In our study, this reference level is always a standard elliptical fit of data points and thus we can reuse the stimuli from task JND again instead of producing replicate stimuli. Figure 5.1 provides a summary of all the stimuli used in the experiment and their re-use for statistical analyses of various task categories.

5.2.1.4 Check points

In addition to the stimuli from the 6 task categories, we have few **check points** to monitor participant's performance during Testing trials and detect *random clicks*. This is done by placing scatter plots having *perfect* positive and negative correlation i.e. $R = 1$ and $R = -1$, respectively at certain intervals during course of the experiment to see whether user estimates them correctly or moves on to the next trial in a

haste. A participant, *random clicking* his way through the trials would fail to observe these relatively *easy* scatter plots and thus his estimate of correlation for these stimuli would be recorded as 0 (which is the default Slider bar value).

Once we have the participant responses, we quickly scan through the data to check values at these trials and see whether participant could estimate them correctly. If not, we deem that user's data unfit for use. In all we have 5 check points at trial no. 15, 35,65 and 85. The initial 3 check point are the stimuli placed on right side in trial no 15, right side on trial no 35 and left side of trial no 65, respectively whereas trial no 85 comprises of the last 2 check points.

Thus, the study has 200 stimuli in total - 10 stimuli for Training sessions, 185 for the Testing sessions and 5 stimuli that act as Check points during the Testing session. Furthermore, we present these stimuli to the users in pairs and thus have only 100 trials in total. We now look at the two primary components of a stimulus design study - data design and scatter plot visualization image, which are discussed in the following subsections.

5.2.2 Data design

In each stimulus, there are a total of 80 data points (except stimuli constructed for task Density category which has 40, 60, 100 or 120 data points as will be explained in Section 5.3). As discussed earlier, each data point in our study is a bivariate data (x and y) and the complete data set for a task category is generated using program written specific to each task category as will be explained in Chapter 6. However, we apply certain rules to our data and data generation techniques, to ensure minimum confounding effect.

Firstly, we avoid negative data point values such that, when plotted, all data points lie only in first quadrant of coordinate plane. To do so, we set the value range on both x and y axis from 1 to 100 and translate all data points of a SCP (obtained via C++ program) by fixed amount. Thus we have a constant scale for each SCP thereby reducing confounding effect due to varying axes scale.

Secondly, some participants may be able to memorize some SCP's data point pattern & cloud shape which are repeatedly shown in accordance with principal of repeated measures. As a result they may always estimate the same amount of correlation for each of them, not based on their perception but past experience of encountering the same SCP. This phenomenon is most probable in case of task JND wherein SCP with correlation 0.7, 0.3, -0.7 & -0.3 occur more frequently. For the very same reason, we generate another alternate set of data for each of the these four correlation values and keep switching between them while presenting to the participants. This is possible because of **many-to-one mapping** between scatter plots and correlation

which says "different scatter plots can correspond to one correlation value but there is only one specific correlation value for each scatter plot". An example is illustrated in Figure 5.5 which shows three different scatter pots each having Pearson coefficient of correlation as 0.

Figure 5.5: Example of many-to-one mapping: each of the 3 SCPs have $R \approx 0$

We refrain from keeping the label IDs of newly formed SCPs similar to previous one to avoid easy identification. Moreover, similar labels tend to *lead* participants towards revealing that the SCPs might have equal correlation. Hence, we provide slightly varied names that are still consistent with the scatter plot labeling scheme as will be discussed in Section 5.2.3.

Lastly, while generating data set for task Progressive symmetry, we ensure that the plotted points for the ellipse and circle do not partially or completely overlap each other which may result in lowering the visual density than what it really is.

5.2.3 Scatter Plot Design

While the data set is generated using C++ program, all the corresponding SCPs are created using MS Excel. Each SCP corresponds to a unique data set and hence renders a different cloud shape corresponding to correlation in data. But they are designed to have similar physical features such as color, shape & size of points plotted and the dimensions of SCPs. We discuss each of them below.

Scatter Plot Labeling The labeling for the SCPs is done in a way that is unintelligible to the participants but conveys to the study designer, the absolute value of correlation along with its direction i.e. positive or negative and the task category that the stimuli belongs to. All this information is incorporated in a 3 to 6 lettered acronym. Table 5.1 specifies the labels of all the scatter plots.

Data point marker

We keep color of the data points plotted on scatter plot to be consistent. We choose a light shade of color blue as it is a neutral and soothing color and complements the white background of the screen. Our initial choices were red and yellow but were discarded considering they are perceived as aggressive and bright, respectively.

All SCPs have the same standard data point shape i.e. circle to eliminate any confounding effect. Each circle has a small size of 2 units to avoid overlapping of

points. We avoid cluttering up the SCP with additional information like axes labels, legends, etc and keep it bare minimum with chart title and the data point cloud only. No labeling for the axes ensures no specific context is provided to any scatter plots and thus we can obtain data-driven estimates only as explained in Section 3.2

Axes Scaling and Tick marks

Both the x and y axes have a fixed scale from 1 to 100 and a spacing after every 10 values, which is displayed in black color. The tick marks have been removed to enhance neatness of the SCP as the users need not perform any value retrieval tasks using them. For the same reason, SCPs have no gridlines for both horizontal and vertical axes which might otherwise obstruct viewing of the data points.

Font

Every SCP has a label and the text used therein uses bold Calibri font of color black and size 14 to ensure consistency.

SCP dimensions

All the SCP have dimensions 10×10 and the two SCPs placed adjacent to each other in a trial are separated by a sufficient distance on screen.

We have covered till now an overview of the stimulus design. The next section will explain in detail each task category and the subsequent stimulus design for each.

5.3 Task specific stimuli design

In the following subsections, we describe the 6 different task categories and the stimulus design for each. We explain the formation of corresponding scatter plot cloud shape by taking into account five control variables namely-

- Distribution Ratio - circle:ellipse:circle(denoted by DR)
- Data Points (denoted by DP)
- Ellipse semi-major axis (denoted by a)
- Ellipse semi-minor axis (denoted by b)
- Circle radius (denoted by r)

Some of them remain constant while others may vary during design of stimulus for different task categories.

5.3.1 Task JND

The aim of this task is to observe whether participants can judge the *existence* of a difference in correlation values between the two SCPs presented in a trial. Next,

we calculate percentage of *correct* response, from all the participant data collected. A response is deemed *correct* if the sign of the difference between actual correlation values of SCPs in a given trial and their estimated correlation values is preserved. For example, given two SCPs with actual correlation values, $R_1 = -0.5$ and $R_2 = 0.9$, a *correct* response would be the one where estimated correlation value for R_1 is smaller compared to R_2.

5.3.1.1 Pairing

Unlike the other task, the stimuli for task JND need to be presented to participants in pairs, adjacent to each other in a trial, to observe whether people can detect a difference in correlation values of the two scatter plots. In all, we have 21 different correlation values ($-1, -0.9, ..., -0.1, 0, 0.1, ..., 0.9$ and 1) each of which can be paired with the others for this task. This means there is a possibility of a total of ${}^{21}C_2 = \frac{21!}{2!(19)!} = 210$ pairs of SCPs. Since its not practically feasible and within scope of the study to present all these pairs to participants, we strategically decide 4 base correlation values, $0.3, 0.7, -0.3$ and -0.7 -against which we measure JND. This is the basis of formation of 4 sub categories of task JND Fine namely, task JND Fine 0.3, task JND Fine 0.7, task, JND Fine -0.3 and task JND Fine -0.7. Table 5.3 shows the pairs of stimuli for each sub task category.

The explanation behind selecting these 4 correlation values as base values is simple. We select a high and a low correlation value each from the positive and negative scale. For the high base correlation (0.7), its partner scatter plot has a comparatively lower correlation (0.6, 0.5, 0.55, etc) while for low base correlation, its partner scatter plot has a higher correlation. The scenario is reversed in case where base correlations are negative. Lastly, the position of the target and variable scatter plot is randomized (i.e. whether the target appears as the left or right).

5.3.1.2 Stimuli

We have 80 unique stimuli for this task, none of which are repeated. They are divided into two sub-categories- JND Coarse and JND Fine each of which has 24 and 56 stimuli respectively. The JND Coarse task category tests pairs of stimuli which have 'even' correlation difference between them such as 0.2, 0.4, 0.6 and 0.8. We present 3 pairs for each of the four values and then average the result when calculating percentage of *correct* response. Table 5.2 contains the list of all stimuli needed and their pairings.

On the other hand, pair of stimuli from JND Fine task category has 'odd' differences between them such as 0.05, 0.1, 0.15, 0.25, 0.3 and 0.35. Ideally, it is expected to have higher percentage of *correct* response for JND Coarse category regardless of the JND whereas in case of JND Fine, this percentage would gradually go up as the

	JND	Pair 1	Pair 2	Pair 3
1	0.2	0.1, 0.3	0.7, 0.9	0.5, 0.3
2	0.4	0.1, 0.5	0.7, 0.3	0.5, 0.9
3	0.6	0.7, 0.1	0.3, 0.9	0.8, 0.2
4	0.8	0.1, 0.9	0.15, 0.95	0.05, 0.85

Table 5.2: Stimuli pairing for task JND Coarse

	Correlation difference in SCPs	JND Fine 0.7	JND Fine 0.3	JND Fine -0.7	JND Fine -0.3
1	0.05	0.7, 0.65	0.3, 0.35	-0.7, -0.65	-0.3, -0.35
2	0.1	0.7, 0.6	0.3, 0.4	-0.7, -0.6	-0.3, -0.4
3	0.15	0.7, 0.55	0.3, 0.45	-0.7, -0.55	-0.3, -0.45
4	0.25	0.7, 0.45	0.3, 0.55	-0.7, -0.45	-0.3, -0.55
5	0.3	0.7, 0.4	0.3, 0.6	-0.7, -0.4	-0.3, -0.6
6	0.35	0.7, 0.35	0.3, 0.65	-0.7, -0.35	-0.3, -0.65

Table 5.3: Stimuli pairing for task JND Fine

correlation difference between the two SCPs increases.

JND Fine is further divided into four sub sub-categories, JND Fine 0.7, JND Fine 0.3, JND Fine -0.3 and JND Fine -0.7 each of which has 14 stimuli each. The only difference amongst these four categories is that they have different reference correlation i.e. R from which the difference is calculated. For example, JND Fine 0.7 has stimuli pairs like (0.7, 0.65) and (0.7,0.35) whereas JND Fine -0.3 has stimuli pairs like (-0.3,-0.55) and (-0.3,-0.4). Also, unlike JND Coarse, there is only one pair of stimuli presented to participants for each difference value. Table 5.3 contains the list of all stimuli needed for all sub task categories and their pairing.

5.3.1.3 Variables

The list of variables that remain constant throughout this task are namely, a, DP, DR and r whereas b and R vary and have an inverse relation between them. The design for all the stimuli remain same regardless of the category they belong to and are subjected to the following specifications:

- $DP = 80$.
- The SCP's cloud shape is elliptical and all the data points lie within it, hence DR=0:80:0.
- $a = 7$.
- The value of b varies depending upon R of the SCP and they have an **inverse relation** between them. Details can be found in Table 5.4.

	Correlation	b
1	-0.7 or 0.7	2.5
2	-0.65 or 0.65	2.6
3	-0.6 or 0.6	2.7
4	-0.55 or 0.55	2.75
5	-0.45 or 0.45	3.25
6	-0.4 or 0.4	3.5
7	-0.35 or 0.35	3.75
8	-0.3 or 0.3	4

Table 5.4: Independent variables and their values for task JND (b is the semi-minor axis of the ellipse

5.3.2 Task Reflective Asymmetry

The aim of this task is to observe how perception of the participants changes with variation of the reflective asymmetry of the cloud shape. The SCP cloud is made of two semi ellipsoids having common major axis (a) but different semi minor axes (b_1 and b_2) (see Figure 4.3).

5.3.2.1 Stimuli

We consider 3 correlation values, $0.9, 0.7$ and 0.5 as key observation points for the task Reflective Asymmetry to analyze its effect on high, low and neutral correlation values. Each correlation value contributes 3 stimuli each and thus we have 9 unique stimuli for this task. Figure 5.6 illustrates stimuli from this task category. These are presented repeatedly (at least 3 times each) to the participants as explained in Section 5.2.1 and the error in estimation of correlation is then averaged for each of the 9 stimuli for each participant.

5.3.2.2 Variables

The list of variables that remain constant throughout this task are namely, a, DP, DR and r whereas b_1, b_2 and R vary. Table 5.5 gives the relation between these 3 independent variables.

The 3 stimuli for each base case (0.5, 0.7, 0.9) are designed with the following specifications:

- DP= 80.
- The distribution of data points in both ellipses is equal and hence DR = 0:80:0.
- a = 7.
- One of the stimuli has b_1=b_2 while remaining two has $b_1 > b_2$.

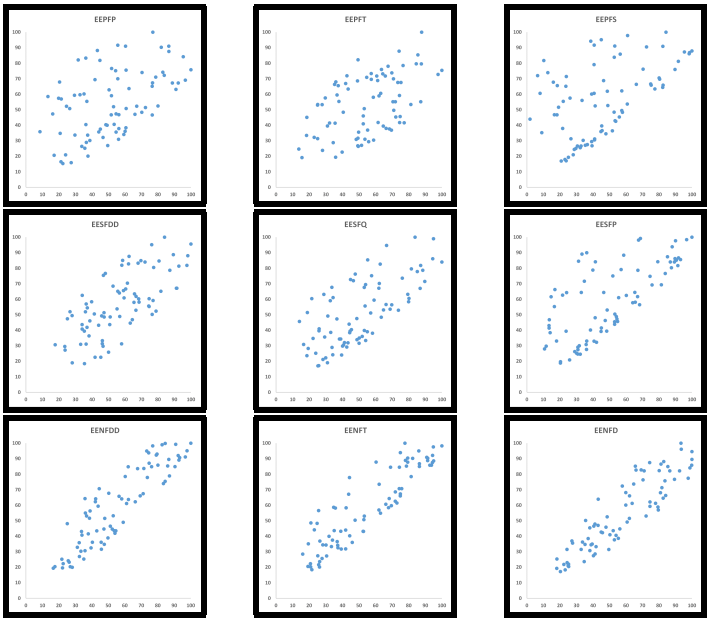


Figure 5.6: Stimuli for task Reflective Asymmetry

Correlation	b_1	b_2
0.5	3	3
	5	2
	7	1
0.7	2.5	2.5
	4	1.5
	5	1
0.9	2	2
	2.5	1.5
	3	1

Table 5.5: Independent variables and their values for Task Reflective Asymmetry where b_1 and b_2 are the semi-minor axes of the two semi-ellipsoids

- r=0.

5.3.3 Task Distribution

The aim of this task is to observe how perception of the participants change with a change in the data distribution of the scatter plots. Each SCP cloud is made of 3 elements, two circles of same radius and one ellipse in between them, amongst which the 80 data points are distributed in varying proportions (see Figure 4.2).

	Distribution Ratio (DR)	Correlation (R)
1	0:80:5	0.5
2	5:70:5	0.25
3	10:60:10	-0.1
4	15:50:15	-0.3
5	20:40:20	-0.5
6	25:30:25	-0.7
7	30:20:30	-0.9

Table 5.6: Independent variables and their values for Task Distribution

5.3.3.1 Stimuli

There are 6 unique and 1 repeated stimuli for this task, as illustrated in Figure 5.7. We consider seven distribution ratios (DR) which gives 7 corresponding correlation values (R) and 7 subsequent stimuli, summarized in Table 5.6.

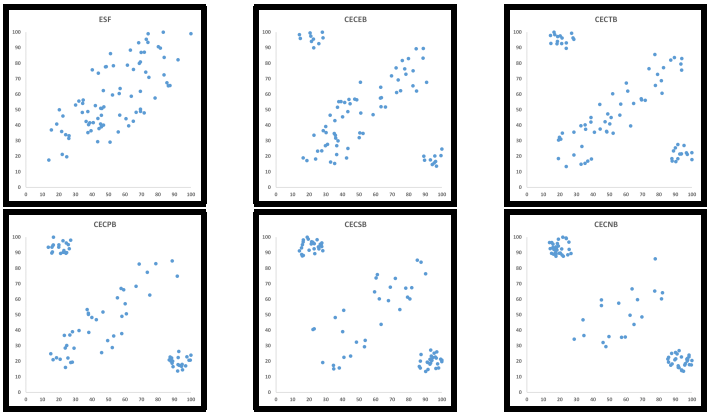


Figure 5.7: Stimuli for task Distribution

Each stimuli corresponding to each of the DR is presented thrice except in case of $DR = 0{:}80{:}0$, which is presented 11 times to each participant. The error in estimation of correlation is then averaged for the repeated measures for each of the 7 stimuli.

5.3.3.2 Variables

The 6 stimuli corresponding to the six distribution ratios are designed with the following specification:

- Each has a unique DR as given in Table 5.6.
- $DP = 80$.
- $a = 7$.

	Semi-minor axis (b)	Radius of circle (R)
1	0.5	5
2	1	4
3	2	3
4	2.5	0
5	3	2
6	3.5	1

Table 5.7: Independent variables and their values for Task Progressive Symmetry

- $b = 3$
- The radii of both the circles are same, $r = 1$.

5.3.4 Task Progressive Symmetry

The aim of this task is to observe how perception of the participants changes with the change in the progressive symmetry of the cloud shape in scatter plot which is made up of an ellipse and a circle at top right corner of ellipse. Figure 5.8 illustrates all the stimuli used in the task.

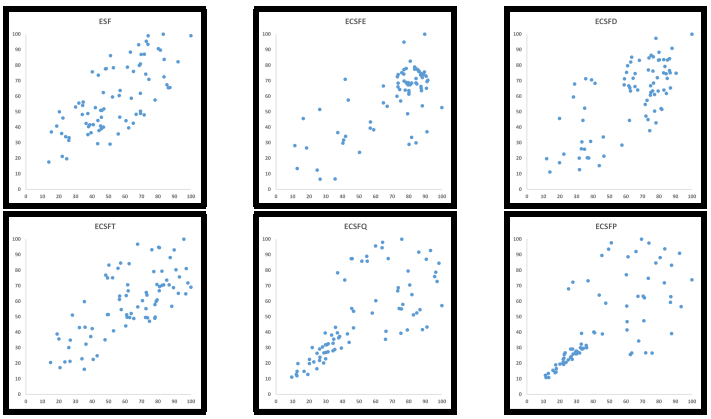


Figure 5.8: Stimuli for task Progressive Symmetry

5.3.4.1 Stimuli

We consider 6 stimuli, each with a different asymmetrical shape, obtained from varying b and r while simultaneously keeping the correlation constant (see Figure 4.4). The 5 unique (b,r) pairs are summarized in Table 5.7:

Each of these stimuli is presented thrice except when b=2.5 in which case it's presented 11 times to each participant, as explained in Section 5.2.1. The error in estimation of correlation is then averaged for the repeated measures for each of the 6 stimuli.

5.3.4.2 Variables

The stimuli corresponding to the six asymmetrical cloud shapes are designed with the following specification:

- Each has a unique (b, r) pair as mentioned above.
- $DP = 80$.
- The data points are distributed equally between the ellipse and the circle, hence $DR = 40{:}40{:}0$.
- $a = 7$.
- Each has a constant correlation, $R = 0.7$.

5.3.5 Task Density

The aim of this task is to observe how perception of participants vary as density of the scatter plots changes. We consider five different densities - 40, 60, 80, 100 and 120 as key reference values. The cloud shape is elliptical for all stimuli in this task category.

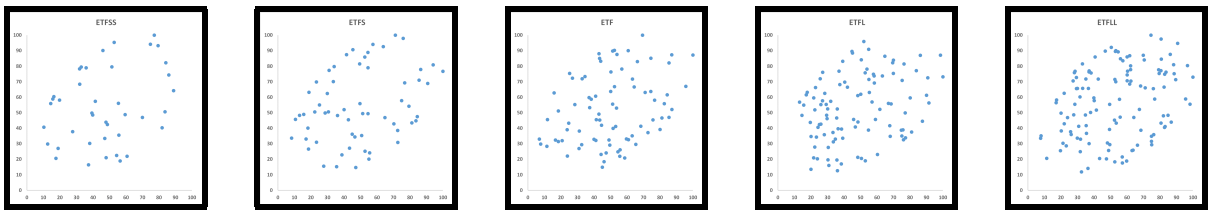
5.3.5.1 Stimuli

For each density, we test participants for 3 different correlation values- 0.3, 0.5 , 0.7. Thus we have 12 unique stimuli for this task, as illustrated in Figure 5.9. These are presented to participants repeatedly (at least 3 times each) during the course of trials. The error in estimation of correlation is then averaged for the repeated measures for each of the 15 stimuli.

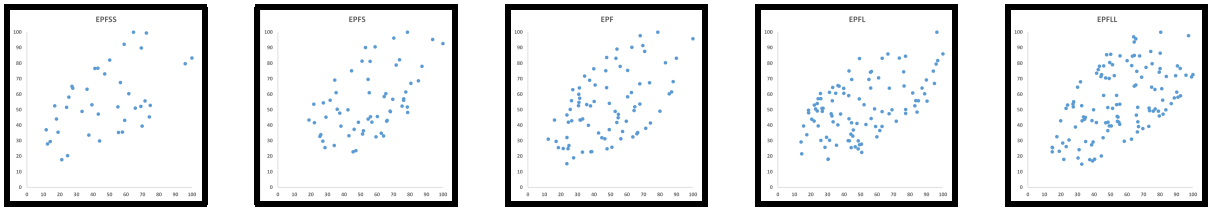
5.3.5.2 Variables

The 15 stimuli corresponding to the five different densities are designed with the following specification:

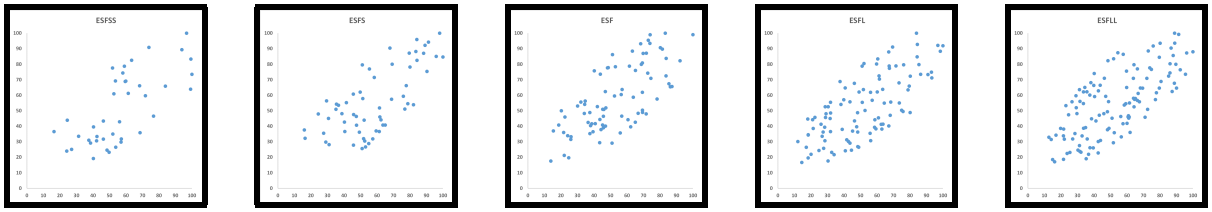
- Each stimuli has a unique (DP, R) pair.
- The SCP's cloud shape is elliptical and all the data points lie within it, hence $DR = 0 : 80 : 0$.
- $a = 7$.
- The value of b depends on R. Table 5.8 describes their relation.
- $r = 0$.



(a) R=0.3



(b) R=0.5



(c) R=0.7

Figure 5.9: Stimuli for task Density for different correlation values of $R = 0.3, 0.5, 0.7$

5.3.6 Task Weber

The aim of this task is to observe the amount of error participants make while estimating correlation in regular ellipse-shaped scatter plots. The shape of the cloud in all SCPs is ellipse, same as in case of task JND.

5.3.6.1 Stimuli

We observe correlations on both positive and negative end for $R = -0.9, -0.7, -0.5, -0.3, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ and 0.9. Sampling at such small intervals improves quality of results and avoid faulty conclusions due to insufficient observations. Each of these scatter plots are presented at least thrice to participants and the error in estimation of correlation is then averaged for the repeated measures for each of the 15 stimuli. The order in which they are presented is randomized as when correlated data is presented sequentially, people fall prey to illusory correlation (Chapman, 1967). Figure 5.10 enlists all the stimuli used.

5.3.6.2 Variables

The design for the stimuli are subjected to the following specification:

- $DP = 80$.

Density	Correlation (R)	Semi minor axis (b)
40	0.3	4
	0.5	3
	0.7	2.5
60	0.3	4
	0.5	3
	0.7	2.5
80	0.3	4
	0.5	3
	0.7	2.5
100	0.3	4
	0.5	3
	0.7	2.5
120	0.3	4
	0.5	3
	0.7	2.5

Table 5.8: Controlled variables and their values for task Density

- DR= 0:80:0
- $a = 7$.
- The value of b varies depending upon R of the SP and they have an **inverse relation** between them. Details can be found in Table 5.9.

Having generated the stimuli for each task category, the next section explains the features and specifications of the software used in the study to capture the user response.

5.4 Software Design

The purpose of the software used in this experiment is to mainly introduce participants to basic statistic concepts of correlation and scatter plots (Training Session) and collect their performance in various correlation estimation task categories (Testing session). The workflow of the software is illustrated in Figure 5.11. It consists of four main parts as can be observed from the diagram, namely Introductory screen (collecting demographic information and familiarity rating), Training trials, Feedback from it, Testing trials, Breaks and Downloading user results.

In the following subsections, first we present an overview of the software, before describing in detail the software workflow, time scheme, and the design of sequence of trials.

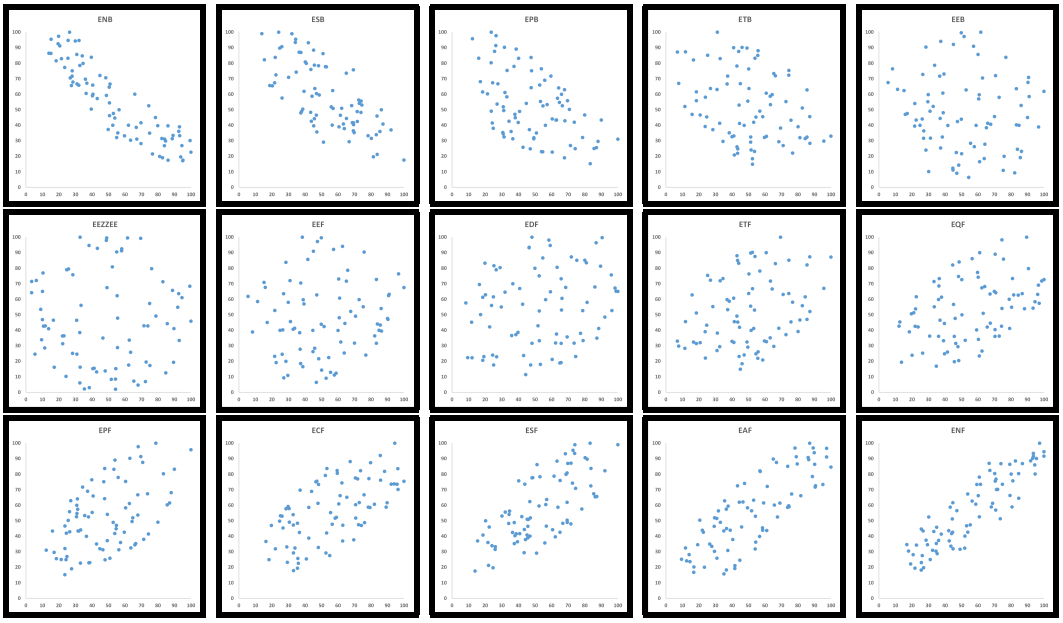


Figure 5.10: Stimuli for task Weber in increasing order of correlation from left to right

5.4.1 Overview

The software implementation is done mainly using HTML, JavaScript and PHP and is used to gather user response for different task categories.

For each trial, 2 SCPs will be shown on screen with a standard question (asking users to estimate correlation for the two scatter plots) and 2 corresponding slider bars in a common row below them (to record correlation value for each (refer to Figure 5.1). The left-right spatial ordering of the two plots was randomly determined meaning it isn't necessary that correlation of left SCP is always greater than that of the right SCP or vice versa. Either of them could be greater or smaller & in some cases both could be equal. The current trial number will be displayed on top the screen. A 'Next' button is placed between the 2 slider bars, which when clicked, will record user response for the 2 stimuli in that particular trial and proceed on to the next trial. At any point, the software disables the user from going back to previous trial to modify his answer once he hits the 'Next' button.

Once the participants has submitted their response for a trial, they are presented with a masking screen for 2 seconds (see Figure 5.14). It is for the sole purpose of introducing some amount of random *visual noise* which is useful for relaxing eye vision between two trials as will be explained in Section 5.4.6.

We refrain from using a countdown timer as we prioritize accuracy rather than efficiency. However, the software will calculate the response time of each user for each trial. This is done by recording *starting time* (as soon as web page is displayed) and *ending time* (when 'Next' button is pressed) and calculating the difference of the two.

	Correlation	b
1	-0.9	2
2	-0.7	2.5
3	-0.5	3
4	-0.3	4
5	-0.1	5.5
6	0	7
7	0.1	5.5
8	0.2	5
9	0.3	4
10	0.4	3.5
11	0.5	3
12	0.6	2.7
13	0.7	2.5
14	0.8	2.25
15	0.9	2

Table 5.9: Controlled variables and their values for Task Weber

This has 2 advantages:

- judge the efficiency and ease of estimation of correlation according to task category.
- identify the system crash point in case of software malfunctioning.

The order of trials in the software is pseudo-randomized in order to minimize the sequence effects but remains same for all the participants, as will be explained in Section 5.4.3. Additionally, the participants are subjected to three short breaks during the trials in order to reduce fatigue and boredom which is discussed further in section 5.4.5.

5.4.2 Software Workflow

In the first part of the software, the participants are required to input their User ID and some demographic information such as age, gender and occupation, using text boxes & radio buttons. They also specify their familiarity rating of SCP using five-level Likert scale which consists of *Not at all Familiar, Slightly Familiar, Somewhat Familiar, Moderately Familiar* and *Extremely Familiar* respectively. Lastly, we record whether the participants have any known type of Color blindness which might hamper their performance during the experiment. Figure 5.12 illustrates this screen.

The second process is the Training session, consisting of five trials. Each trial is designed to familiarize the participants with the 'look and feel' of the software on a

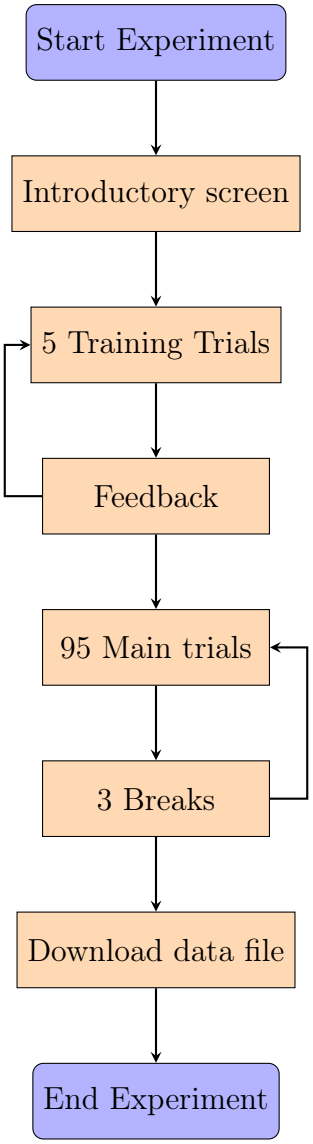


Figure 5.11: Software workflow

whole and acquaint them with instructions on how to perform the task accurately. To facilitate learning in each trial, the participants are given feedback in form of correct answers, which they can then compare with their own answers and improve incrementally. To enable this, the Training trials have a 'Check my Answer' button (in place of 'Next' button), between the two slider bars which when clicked, shows the mathematical formula-calculated correlation values for each scatter plot in a pop-up alert box. All trials are strategically placed beginning from easiest to toughest as explained in Section 5.2.1.2.

The third step of the workflow is the main Trial session. In this session, a total of 95 trials are presented to each participant. However, unlike the Training session, no feedback is provided.

Together with the 5 Training trials, we have 100 trials which are conducted in 4 batches of 25 trials each. There is a 2 minute, 4 minute and 2 minute break after each

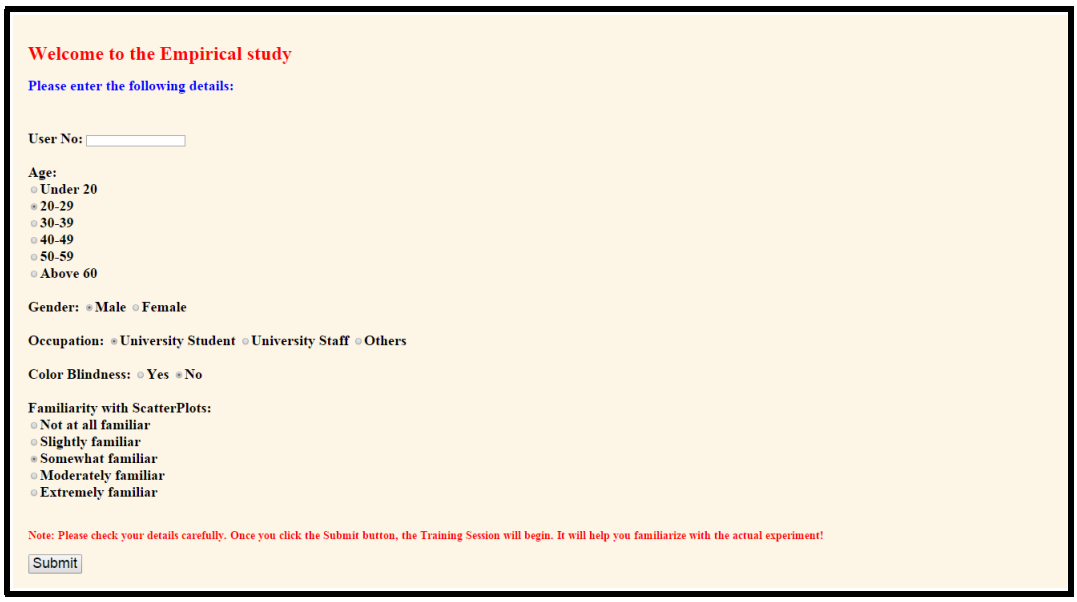


Figure 5.12: Introductory screen of the software used to collect participant data

batch, as will be explained in Section 5.4.5

In the last part of the software, the participants are required to download the file containing their responses for each trial and email it to a backup account. This is done so as to maintain backup copies in case the original data is lost in the unlikely event of hard-drive failure, system crash, virus attack or natural disaster.

5.4.3 Sequence Design

As explained in Section 2.5.3, stimuli provided in the earlier trial may create positive or negative impact on user's judgment of the stimuli displayed later. In order to reduce it, we have provision of a masking screen for 2 seconds between each trial which reduces the confounding effect.

Furthermore, the user's judgment of a stimuli may also be influenced by the second stimuli present next to it on the screen. We could not place individual stimuli in a trial as some task like JND requires stimuli to be presented in pairs. We also couldn't present some stimuli in pairs while others as standalone stimuli as it would lead to *leading effect* towards Task JND. Hence we make a tradeoff and present them all in pairs. However, to minimize the effect of the adjacent stimuli, we ensure all the 3 copies of a stimuli (generated in accordance with principal of repeated measure) from a task category are paired with different stimuli each time. This pseudo-randomization scheme also ensures that the stimuli from the same task category are separated by a distance of at least 6 trials. In addition to this, we also randomize the left-right spatial ordering of the two scatter plots on the screen so that the participant is unable to associate the positioning of the scatter plot on the left or right side of the screen

with a higher or lower correlation value.

5.4.4 Time scheme

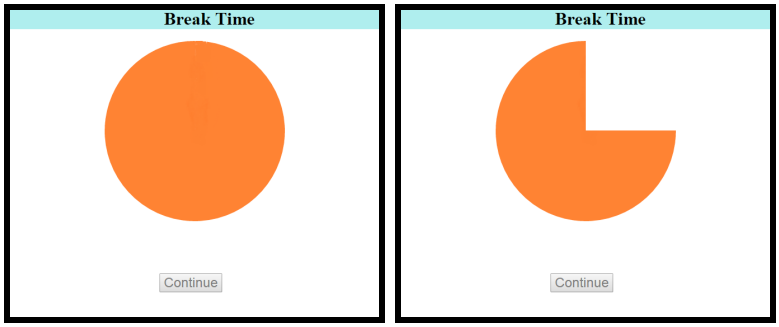
The software has in built functionality for calculating user's response time per trial. However, our study does not place a limit on time taken to make estimations for a trial, due to two main reasons. Firstly, with time limitation, if the participant is unable to make estimations in stipulated time, he may choose to guess and randomly scroll the slider bar in either direction and thus the result might not be correct. Secondly, some participants may use a long time to do some difficult trials compared to others and thus putting a bound on time limit won't be justified. Our ultimate aim is to obtain accurate results rather than efficient results. For these reasons, we do not set a time limitation for each trial.

5.4.5 Break scheme

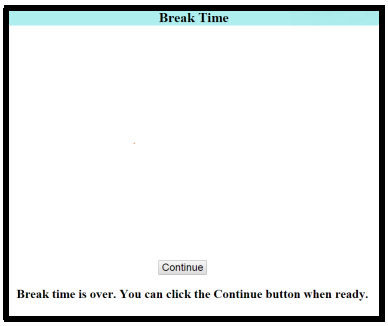
Some participants may experience fatigue and boredom after repeated trials, which may affect their performance in later trials. On the other hand, some participants might not be tired but they might develop some learning curve and memorize cloud shapes after repeated viewing of the SCPs. Therefore, we subject all the participants to three short break of *at least* 2 minutes, 4 minutes and 2 minutes during the course of the study. This helps them ease the tiredness before continuing to the next set of trials. Our study does not limit the time for the break as different participants may require different times to recover from fatigue.

During the break, the participants are presented with a *disabled* 'Continue' button & a pie clock on the screen in front of them which reports the amount of time left for the break to be over. Once the break time has elapsed and the clock has completely vanished from screen, the 'Continue' button on the screen is enabled and an instructions appears indicating that the participant, if ready, may proceed to further trials. The user may then click on it and continue with the next batch of trials or may choose to relax a bit longer. Figure 5.13 illustrates the screen a) at start of the break, b) during the break and c) at the end of the break.

The amount of time assigned for the breaks has been chosen intuitively. During the pilot study conducted with two participants - 1 female and 1 male, the average time to complete the 100 trials (sans breaks) came out to be 32 minutes. We assign approximately 20 minutes to deliver the pre-study presentation, which leaves an 8 minute window for breaks, given that we have set 60 minutes as target time to complete the study. The participants are expected to be least tired at start of study, hence the first break accounts for 2 minute only. As the trials progress, they may



(a) at start of the break (b) during the break



(c) at the end of the break

Figure 5.13: Break screen in the software at different time intervals

encounter fatigue and hence a comparatively longer break of 4 minutes is provided. Accordingly, the last break is assigned a shorter duration of 2 minutes.

5.4.6 Masking screen

A number of recent behavioral and neuroimaging studies suggest that, at least under some circumstances, increasing attention to one task can enhance performance in a second task (e.g., the attentional boost effect) (Swallow and Jiang, 2013) while others suggest that as attention to one task increases, performance in the second task always suffers (Kinchla, 1992). On the other hand, some studies suggest that increasing attention can have both impairing/negative & improving/positive impact depending upon the spatial resolution of the image(Yeshurun and Carrasco, 1998).

Nonetheless, what is central to all these studies is the fact that continuous exposure to these trials might impact user performance in later trials. Thus, in order to keep the user performance consistent, we introduce a visual noise screen between two trials, during both the training and testing sessions. Participants are subjected to a blank screen as shown in Figure 5.14. It is displayed for 2 second only, following which the next trial occurs.

In field of psychology this is known as **Masking effect** wherein one brief stimulus,reduces or eliminates the visibility of the main stimulus for a certain duration. Out of the two Masking types, we incorporate 'Backward Masking' in our study which

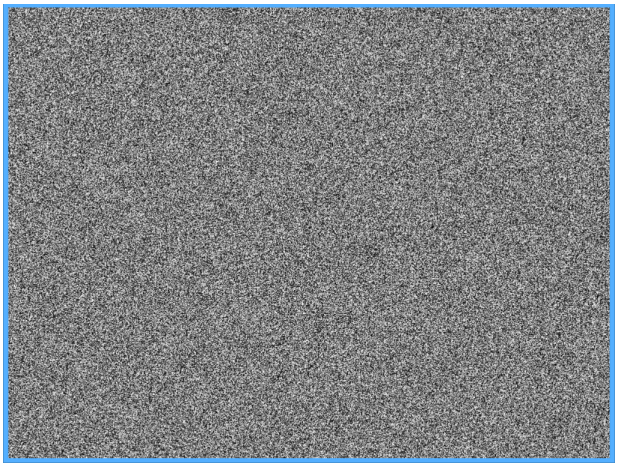


Figure 5.14: Masking screen

means the target stimuli precedes the masking stimuli. Contrary to what intuition might suggest, 'Backward Masking' has its strongest influence not when both target and mask objects are present on the same screen but when a brief temporal gap is presented between the two (Enns and Di Lollo, 2000). Hence, this ensures that participant's estimate for a particular stimulus is not influenced by any of the previous trial's stimulus.

5.5 Pre-study presentation

To acquaint all participants with basic Statistic knowledge required to perform the experiment, we provide them with a 20 minute power point presentation. The need of such a presentation and its contents are explained in the following sections.

5.5.1 Overview

The aim of an empirical study is to facilitate the research based on actual experimentation and observation by end users rather than merely relying on theory or belief. In order to enhance the user experience and ease their interaction with the whole study, we present a brief and informative pre-study PowerPoint presentation to them. Also, since we draw participants, from various academic backgrounds(Arts, Science, Mathematics, Philosophy, Law, Economics, Computer Science, Engineering etc), it is important to theoretically bring all of them at par with each other in terms of understanding of basic statistical concepts (Correlation, Scatter Plots, etc) used throughout the scope of the study as seasoned analyst are likely to interpret the scatter plot differently than novices. It is important to control participant's background knowledge and familiarize them with prerequisite knowledge about the topic so as to avoid confounding effect due to different knowledge domains which in turn may render

biased user result. Since the topics are conceptually not very difficult to comprehend, we incorporate them in a short interactive 20 minute presentation.

5.5.2 Contents

It covers the aim of the study, the participants' role for the study and what we as researchers are trying to determine via means of this study. This is done so that they are apprehensive of their contribution to the visualization community and they perform the study with utmost sincerity and perseverance.

We presentation begins with explanation of correlation and scatter plots using real life examples. Next, it depicts in a flowchart, their itinerary for the day which includes a Training & Testing session done in four batches with breaks in between, emailing the backup file & filling feedback form. It also contains screenshots of Training and Testing session trial screens to familiarize the participants with it and explain the working of the software. Lastly, certain instructions related to software use are presented.

They are reminded orally, not to make random guesses and cautioned against assuming that a particular range of correlation values was necessarily covered or that the correlation values are necessarily ordered. Any doubts and general queries from participants are answered during the delivery of the presentation.

The presentation slides are provided in Appendix A for reference purpose.

5.6 Feedback Survey Design

Once the study has ended, the participants fill a paper-based feedback survey form in which they are required to provide the 'ease-of-estimation' rating for each task category. It has three multiple choice questions and the rating options for each uses a modified version of five-level Likert scale, similar to one used for familiarity rating in the introduction page of software. Each question specifies two of the contradicting task categories and a generic question which asks users to select the easier (much easier) of the two.

The three questions compare ease of estimation in case of 1) positive & negative scatter plots, 2) uniform & non uniform scatter plots and 3) high density & low density scatter plots respectively. The participants can choose from options such as *Category 1: much easier, Category 1: easier, same, Category 2: easier* and *Category 2: much easier*. There are three stimuli presented for each category for users to base their comparison on and select an appropriate answer. All of them are identical in dimensions and other SP attributes (axes scale, color, etc) to eliminate any bias. The only difference, compared to the ones presented during the study is that the axes scales have spacing after every 20 values rather than 10 due to the small size of the

stimulus in feedback form.

The feedback forms are distributed at the end of the study rather than in the beginning itself to avoid *Experimenter Effects* where the experimenter unconsciously conveys to participants how they should behave (this is called experimenter bias). The feedback form questions might give unintentional clues to the participants about what the experiment is trying to investigate and how they expect them to behave. This affects the participants' behavior.

The feedback form has been provided in Appendix B for reference purpose.

Chapter 6

Implementation

The implementation process consists of three main components- stimulus generation, software implementation and the experiment itself.

Firstly, we perform stimulus generation and then proceed to software implementation, both of which are done iteratively using Agile software development, which assures software quality. Finally we conduct a controlled experiment using the developed software to collect user performance on the correlation estimation task using scatter plots.

In totality, there are total of 3500 trials included in our study (including training trials): 35 participants $\times$100 trials (including task JND, task Weber, task Reflective Asymmetry, task Progressive Symmetry, task Density, and task Distribution).

We begin by describing the software development process in Section 6.1. In the second section, emphasis is on details of execution of the experiment.

6.1 Software Development Process

Since ours is a time critical project, we deploy the use of Agile Software Development (ASD) process for stimuli generation and software implementation. It is a type of "incremental model" wherein the software is developed in incremental, rapid cycles with each incremental release building on previous functionality. Each release is thoroughly tested to ensure software quality is maintained. This 'incremental' and 'iterative' approach helps develop software at the same time we're gathering requirements. ASD can better handle requirement changes from stakeholders, which comprises of our research team members and generate an executable product at end of each cycle. Feedback provided in one cycle is worked upon for the next release, which is produced in a short time frame. In our project, we produce each increment in a one-week period.

Because the software and the stimuli can be developed independently to each

other, we separate the development process of the two to concentrate on only one at a particular time. The following subsections describe the iterations involved in both the development procedures.

6.1.1 Stimuli Generation

We began with stimuli generation, which took approximately 8 weeks for completion.

Iteration 1

- Devise 'secret' label ids for SCPs as explained in previous chapter in Section 4.2.3.
- Design and generate data sets for Task JND.
- Generate corresponding SCPs.

Iteration 2

- Define axes scaling and modify data set for Task JND.
- Specify dimensions (height & width) of SCP to ensure all SCPs are uniform.

Iteration 3

- Design and generate data set for task Distribution and Reflective Asymmetry.
- Generate corresponding scatter plots.

Iteration 4

- Modify metrics (a, b, r) used to generate SCP for Task Distribution.
- Generate second version of data set and corresponding SCP for Task Distribution.

Iteration 5

- In Task JND, for all repeated correlation values, generate an additional data set.
- Design the pairing up of stimuli for Task JND.
- Design and generate data sets for Task Density and Task Weber.

Iteration 6

- Model the design of cloud shape of SCPs for Task Progressive Symmetry.
- Generate data set and corresponding SCPs for Task Progressive Symmetry.
- Assign chart titles (secret labels) to each SCP.

6.1.2 Software Implementation

Once we generated all the stimuli relevant for each task, we began with the software implementation which took 5 weeks for completion.

Iteration 1

- Design software workflow & develop software for introductory questionnaire.
- Implement software for displaying a customized slider bar that captures user input.

Iteration 2

- Design stylesheets to enhance layout of document.
- Develop software for main trial session.
- Design timing procedure.

Iteration 3

- Implement functionality for displaying 'masking screen' for 5 seconds between 2 trials.
- Formulate information to be stored for each user record.

Iteration 4

- Develop software to store user results on the server.
- Update duration of masking screen to 2 seconds.
- Change interface for familiarity rating to use Likert scale.
- Change interface for age to use range rather than absolute number.

Iteration 5

- Implement functionality in the software to store user results locally on the system.
- Develop software for training session.
- Introduce break screen between the main session trials.

Iteration 6

- Improve software to enhance user interface & usability and ensure all needed results are recorded.

- Prepare pre-study presentation.

The design for tasks, stimuli and software are explained in Chapter 4, whereas software implementation pseuocode is provided in Appendix D and stimuli generation code is provided in Appendix E for reference.

6.2 Experiment

The experiment was conducted in 7 sessions, 1 for the pilot study and other 6 for the real experiments. All the sessions took place at the Department of Computer Science, University of Oxford.

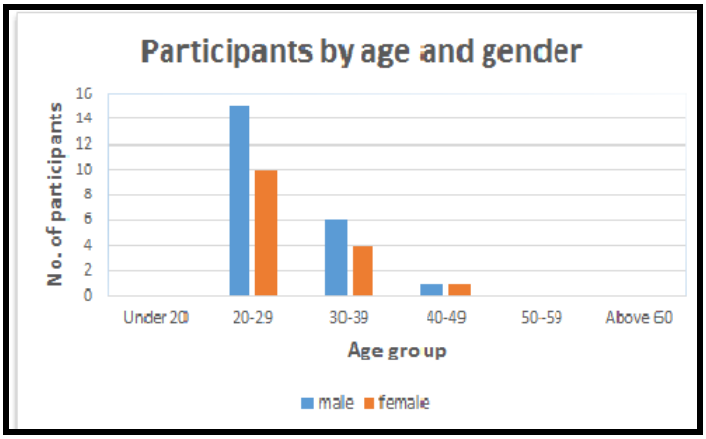
We conducted the pilot study with 2 participants - 1 male and 1 female, who performed the experiment consisting of 100 trials without any rest breaks. The main aim of the pilot study was to evaluate our study design and its implementation. We also measured the average time taken to complete the experiment of 100 trials at a stretch. It provided an estimate of how much time could be allocated for breaks, given that we had fixed an upper limit of 60 minutes for the complete experiment including the pre-study presentation. The pilot study went smoothly and no potential problems were discovered.

The feedback from the pilot study participants suggested fatigue and boredom after initial 25 trials itself. This was the motivation behind introducing 3 short breaks during 100 trials.

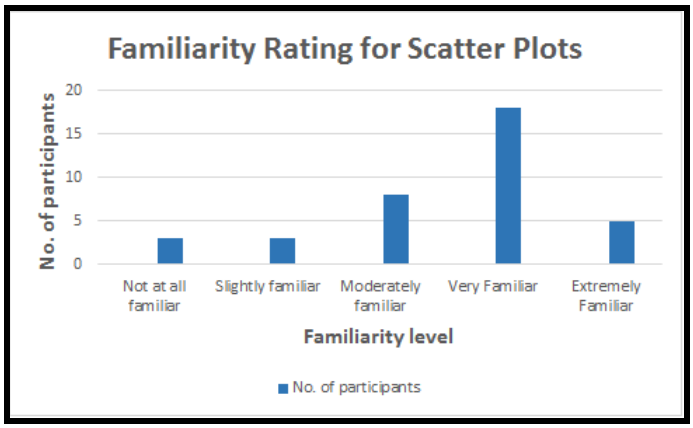
For each session during the main experiment, we limit the number of participants between 4 and 10 to give adequate attention to each. In the following subsections, we give details of the participants who took the study, the apparatus used during the experiment and the procedure that followed.

6.2.1 Participants

A total of 37 participants took part in the experiment in return for a £10 Amazon gift voucher. Among these, there were 27 males and 10 females. Most of the participants comprised of University staff and students, from varied academic background such as Zoology, Engineering, Education, Chemistry, Medical Sciences, Law, etc. A diverse participant population ensured unbiased analyses results. Out of 37 participants, 10 belonged to 30-39 age group, 2 belonged to age-group 40-49 and remaining 25 are in the 20-29 age group. Figure 6.1 illustrates the participant's demographics information and familiarity ratings.



(a) Demographics information including age and gender



(b) Participants' familiarity rating of Scatter Plots

Figure 6.1: Demographics of age and familiarity rating

6.2.2 Apparatus

The pre-study presentation was given on a 10x12 inch projector prior to starting the experiments. The experiments were run on computers with 3.7 GB RAM, 3.30 GHz quad-core Intel core i5-3550 processors running on Fedora, a Linux based OS with GNOME version 3.4.2. Each computer had 24 inch Dell's LCD display with 1920x1200 resolution and sRGB color mode display. We adjust all monitor screens to the same brightness and level of contrasts and set the sleep time of the system to 'Never'. Each participant was required to interact with the software using the mouse and/or keyboard on the desk. A total of 6 to 10 computers were used during a session and all of them were connected to a PHP server. Although the software is cross-browser compatible, to maintain uniformity in display and text font style, we setup the experiment on each system using Chrome web browser in full screen mode. The full screen mode has dual benefits: 1) It disallows users from quitting the experiment mid-way and 2) allows full concentration in software environment as no background distractions are present.

6.2.3 Procedure

The entire study was completed in approximately 40 to 60 minutes, including 20 minutes of pre-study presentation. The time spent on the experiment varied according to the time taken by participants to perform each trial and have a break.

Prior to the experiment, the experimenter presented a brief introductory PowerPoint presentation to the participants. It allowed them to grasp basic understanding of statistical concepts used in the study and familiarize them to the usage of the software.The presentation first explained correlation, its types & properties using 4 real-life examples & corresponding pictorial data tables and scatter plots. Next the screenshots of the actual software the participants have to work upon, were presented to them. This helps them understand the software better and accustom themselves with various elements on screen like slider bars, SCPs, trial no, 'Next' button, etc. Lastly the participants were allotted the UserIDs and some general instructions regarding the trials were provided. Participants were requested to refrain from 'random clicking' and disturbing others during the study.

Next, information sheets were distributed which the participants ought to sign, to imply that they agree to take part in the study and are aware of the aim of the study, risk associated with it & the nature of their involvement. Henceforth the participants are asked/allowed/requested to proceed with the study using the software.

The participants must first provide their user ID, demographic information (age group, gender, occupation, color-blindness) and familiarity rating in the software program. Then, they undertake a training session with 5 trials. Participants are encouraged to clarify their doubts, if any, during the presentation or the Training trials only, else If done during the main trials, the *response time* value might be affected.

After the participants completed the training trials, they were presented with a screen informing them of the end of the training session and a Continue button, to be clicked when the participant were ready to proceed to the main experiment. The entire experiment consisted of 100 trials in total which were presented to participants in 4 batches of 25 trials each. After each batch, the participants were subjected to a short break and were presented with a screen with a pie clock in the center indicating the amount of break time elapsed. The participants could continue to next batch once break time is over and they are ready.

When all the trials were completed the participants were presented with a screen indicating completion of the experiment and a button to be clicked to terminate the software. Upon clicking the button, a file is downloaded automatically containing the user response which were then emailed to a backup account. Once that's done, the participants are acknowledged for their time and service using a £10 Amazon gift voucher.

Figure 6.2 illustrates the entire procedure in a flowchart form.

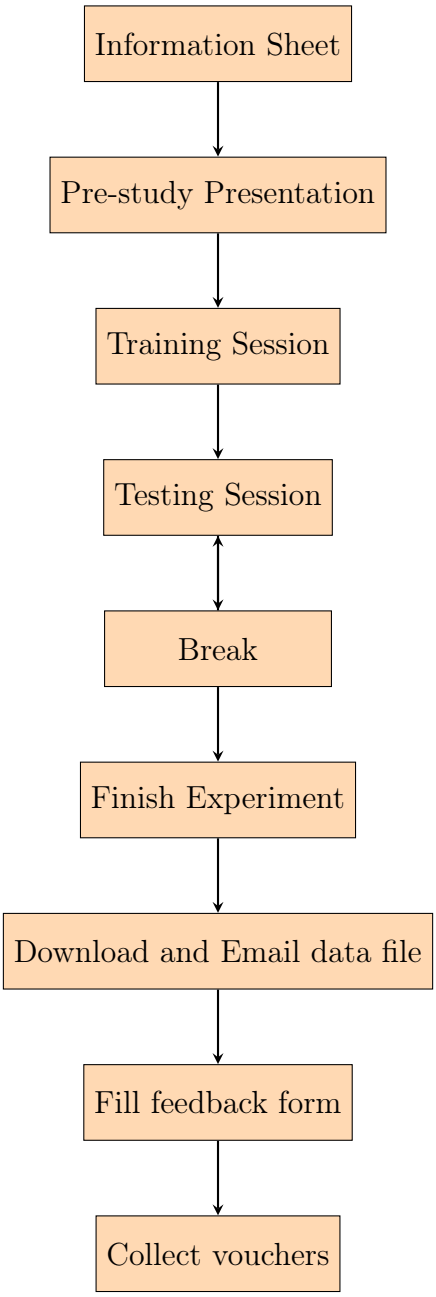


Figure 6.2: Workflow of the experiment

Chapter 7

Result Analyses

As part of the descriptive analysis, we calculate the mean, range and standard deviation for each result. Inferential analysis will be directed towards resolving the hypothesis testing and concluding valid inferences from the study results. In this chapter, we first give a brief account of the validation of participant data. We summarize the cumulative result in Section 6.2. We then move on to the detailed analysis of the results for each of the six tasks separately in Section 6.3 to Section 6.8, respectively. Section 6.9 deals with the subjective analysis based on participant feedback after the experiment.

7.1 Validation of Participant Data

As explained in Section 5.2.1.3, we install several *check points* in our experiment to detect people whose lack the basic essential understanding of estimating correlation using scatter plots, in general and who *random click* for most part of the experiment. This helps filter out data that is unsuitable for analysis as it may render biased results.

A total of 37 participants took part in our study, out of which, we had to remove the data of 2 participants because of the following reasons:

- The estimates for most of the Training trials had an absolute error of more than 1.7 while the acceptable limit is 0.3. The participant failed to estimate correlation correctly even for the *easy* SCPs having correlation as $+1$ and -1 respectively .
- The participant's data consisted mostly of 0s which indicated random clicks. As a result, the participant failed to estimate the values correctly at multiple *check points* too which consisted of *easy* scatter plots having correlation values as $+1$ and -1.

Thus, we had a total of 35 participant's data for analysis. Among these, there were 27 males and 8 females. We next look at task wise analysis of results to test the respective hypothesis.

7.2 Result Summary

We gather the following results *(R1, R2, R3, R4, R5 and R6)* from the analysis of the data.

R1: The JND in correlation is dependent on the reference correlation point chosen and lies between the range 0.05 and 0.1.

R2: The human perception of accuracy of correlation estimation in scatter plots cannot be modeled using Weber's law.

R3: The human perception of correlation in scatter plots is affected by change in Distribution.

R4: The human perception of correlation in scatter plots is affected by change in Density.

R5: The human perception of correlation in scatter plots is affected by change in Reflective Asymmetry.

R6: The human perception of correlation in scatter plots is affected by change in Progressive Symmetry.

Following these, we are able to confirm all of the postulated hypothesis H_1, H_2, H_3, H_4, H_5 and H_6 in Chapter 4.

1. The descriptive analysis of task JND reveals that the JND value of correlation in a scatter plot is dependent on the initial reference point chosen and is most likely to lie in the range $0.05<\text{JND}<0.10$.

2. The descriptive analysis of the task Weber showed no statistical difference between estimation of positive correlation in relation to its negative counter for odd correlation values such as $R = \pm 0.1, \pm 0.3, \pm 0.5, \pm 0.9$, except for $r = \pm 0.7$. In that case, participant performance is better in the positive direction compared to the negative direction.

3. The analysis of task Distribution reveals that a small distribution level does not effect the user perception but subsequently as the distribution level keeps on increasing, the accuracy in user performance decreases.

4. The analysis of task Density reveals lower density values such as 40 data points, renders a significant difference in user performance. However, the user performance is consistent at larger density values such as $60, 80, 100$ and 120. Also,

the user perception is the least accurate at larger correlation values such as $R = 0.7$.

5. The analysis of task Reflective Asymmetry revealed there is no effect of variation in reflective asymmetry for scatter plots for lower correlation values such as $R = 0.5$. However, performance at higher correlation of $R = 0.7$ were affected and further more for $R = 0.9$, thereby concluding that as the correlation increases, people's perception is negatively influenced by the presence of reflective asymmetry and thus their performance accuracy decreases.

6. The analysis of task Progressive Symmetry reveals that the use performance is affected by the variation in progressive symmetry. The reduction in accuracy is directly proportional to the extent of progressive symmetry, irrespective of the direction i.e. more the variation (increase or decrease in ratio of the two semi minor axis), more is the decease in user performance.

7.3 Result Analysis for JND Task

The task JND observed the variation in "precision" of people's perception when they were asked to identify, out of two scatter plots, the one with greater amount of correlation. Here, "precision" refers to the ability of humans to detect a variation in correlation coefficient between two scatter plots. The task was further divided into two sub-tasks namely, JND Coarse and JND Fine. The stimuli pairs for each of the sub tasks were explained in Chapter 5.

7.3.1 Graphical Analysis

Figure B.1 displays the percentage of the participants who could correctly estimate a particular amount of difference between two scatter plots (henceforth referred to as d) for task JND Coarse. As expected, a considerably larger population percentage can detect smaller differences of 0.2 and 0.4 whereas the entire population could perfectly estimate correlation differences of 0.6 and 0.8 between two scatter plots.

Figure 7.2 displays the participant performance in the task JND Fine who could correctly estimate d, with reference correlations -0.3 and 0.3. The two graphs represents similar pattern of correct response percentage. In both the graphs, the percentage increases as the difference between two scatter plot increases since becomes easier to make distinction between the two of them. Also, a sharp increase in the percentage is observed as d increases from 0.05 to 0.1 suggesting that the JND lies approximately in the range 0.05 to 0.1. The same observations hold true for the two

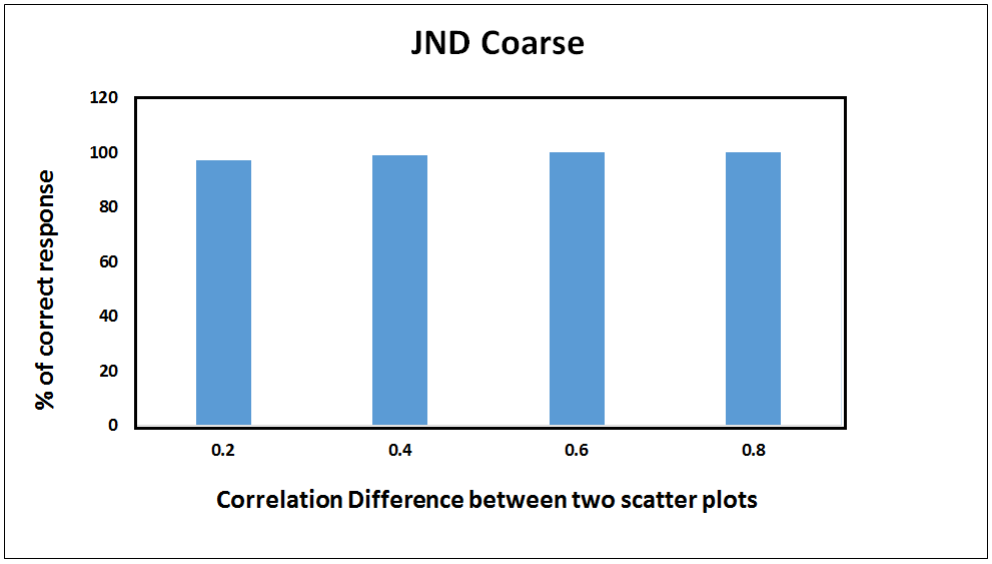
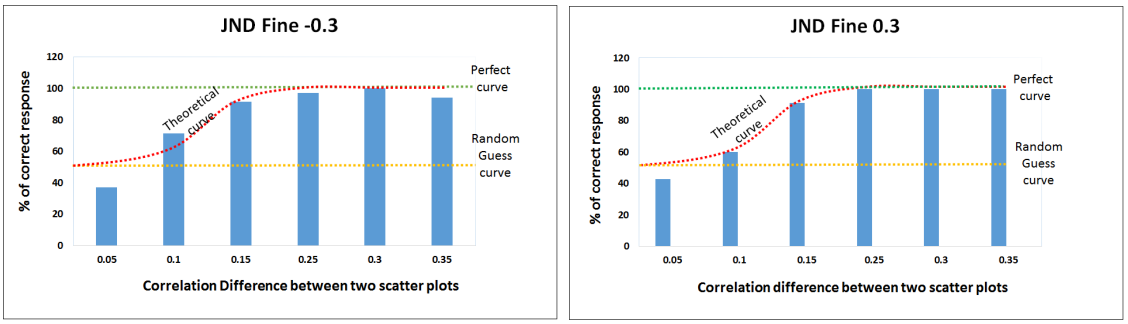


Figure 7.1: Performance analysis for task JND Coarse

graphs in Figure 7.3 which represent the task JND Fine with -0.7 and 0.7 as reference correlation values, respectively. Looking at all the four charts altogether, we can safely conclude that the JND for correlation estimation is not independent of the reference correlation chosen.

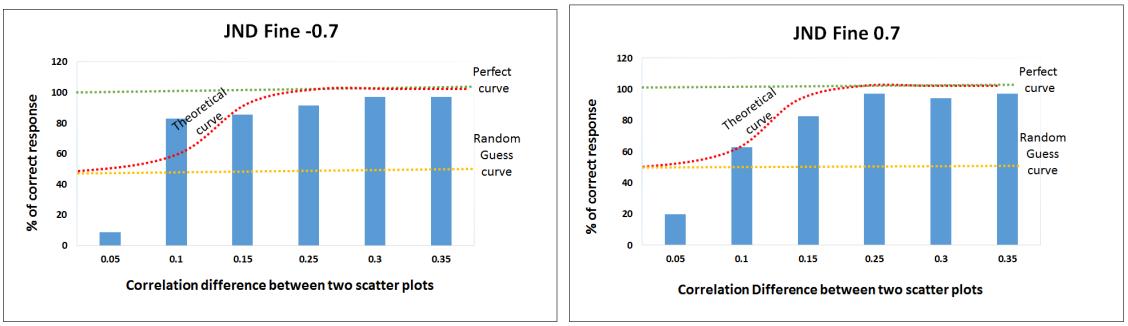


(a) JND Fine -0.3

(b) JND Fine 0.3

Figure 7.2: Performance Analysis for Task JND Fine with reference correlation -0.3 and 0.3

In all the four graphs, the three dotted lines in yellow, red and green give additional information. The upper line (green), called the "perfect curve", depicts the ideal shape of a JND curve when the participant guesses d perfectly every time, which is rarely ever the case. The lower line (yellow), called the "random guess curve" depicts the expected curve shape of the JND graph in the case when participant guesses randomly and hence has a 50% of being correct, which is reasonable enough. The dotted red curve shows the "theoretical curve" which depicts how in reality the JND curve looks like, i.e. gradually increasing as d increases and then becomes the asymptote of the "perfect curve" at higher values of d. However, in all the graphs we observe, at $d <= 0.05$, the percentage of correct response lies further below the "random guess



(a) JND Fine -0.7 (b) JND Fine 0.7

Figure 7.3: Performance Analysis for Task JND Fine with reference correlation -0.7 and 0.7

curve" which suggests that even the random guess is biased suggesting that it is difficult to make distinction between the two scatter plots. Further, all the graphs collectively suggest that the minimum d, required to be reliably discriminated 50% of the times lies between 0.05 and 0.1.

Figure 7.4 illustrates a bubble chart which depicts the cumulative results from the five graphs (one for JND Coarse and four for JND Fine). We refrain from plotting $d > 0.35$ as they are most likely to be estimated correctly by the entire population. The area of the bubble denotes the number of people who correctly detected d at a particular percentage level of correctness. A 100% correctness level means the participant correctly detected d every single time it was presented to him whereas a 0% correctness level depicts the participant failed to detect the difference even once. For example, the bubble in orange color depicts that 6 participants (out of a total 35), correctly guessed $d = 0.15$, 50% of the times'.

As is observed from the bubble plot, a larger proportion of bigger bubbles are concentrated in the 100% correctness level category for larger values of d. For smaller values of d, the big bubbles are concentrated near the lower percentages of correctness level as only a small participant population could correctly detect lower values of d. The observations are in agreement with our previous observations that the accuracy of user estimation significantly drops as d decreases from 0.15 to 0.1 and further more from 0.1 to 0.05.

7.3.2 Friedman Analysis

We perform the Friedman analysis to support our graphical observations, which suggests that there is no statistically significant difference between user performance for detecting *coarse* correlation differences of $0.2, 0.4, 0.6$ or 0.8, $(\chi^2(3) = 3.667, p = 0.300)$. Table 7.1 provides the descriptive analysis of the user performance for task JND Coarse. As the Friedman Ranking is based on the *number of times a correct*

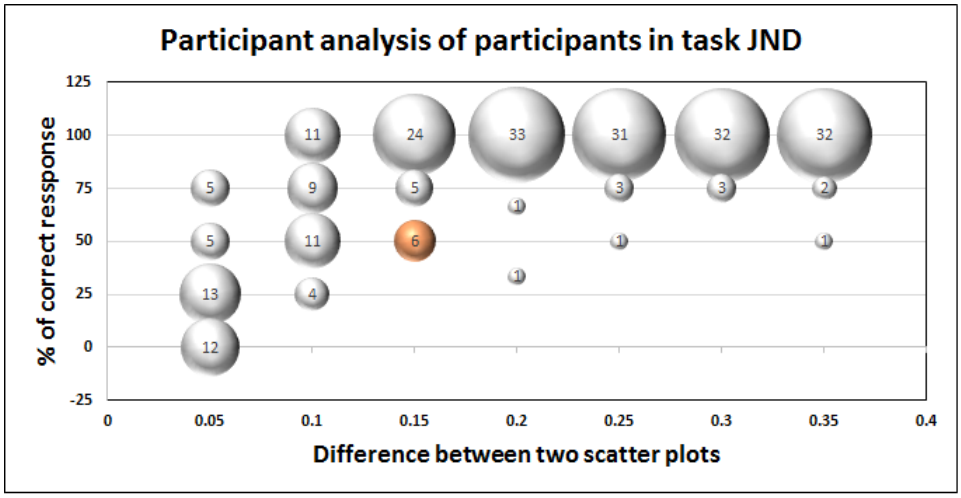


Figure 7.4: Bubble chart representing the population percentage which correctly estimated a difference d between two scatter plots. The bubble in orange is an example signifying 6 people correctly guessed a correlation difference of $d = 0.15$, between two scatter plots, 50% of the times.

estimate of correlation difference is made rather than the difference of actual and estimated correlation, a higher rank suggests better performance. Accordingly, we observe a better user estimation of larger values of d such as 0.6 and 0.8 which rank the highest followed by 0.4 and lastly 0.2.

Descriptive Statistics								
Correlation difference (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.2	2.43	0.97	0.12	0.33	1.00	1.00	1.00	1.00
0.4	2.49	0.99	0.06	0.67	1.00	1.00	1.00	1.00
0.6	2.54	1.00	0.00	1.00	1.00	1.00	1.00	1.00
0.8	2.54	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Table 7.1: Descriptive Analysis for task JND Coarse showing Mean Rank, Minimum, Maximum and Percentiles

Table 7.5, Table 7.4, Table 7.3 and Table 7.2 gives the descriptive statistics for the task JND Fine at the four reference correlation values and suggests that there is a statistically significant difference between the user performance while detecting d when the reference correlation point (RP) varies. For $RP = -0.7$ ($\chi^2(5) = 102.692, p = 0.000$), $RP = 0.7$ ($\chi^2(5) = 86.385, p = 0.000$), $RP = -0.3$ ($\chi^2(5) = 65.366, p = 0.000$) and $RP = 0.3$ ($\chi^2(5) = 72.935, p = 0.000$). From all the four tables, we observe that the participant performance is most accurate when $d = 0.35$ and least accurate when $d = 0.05$.

We further analyze which of the four value of RP best facilitate detection of each of the six values of d. Table 7.6, Table 7.7, Table 7.8, Table 7.9, Table 7.10 and Table 7.11 show the descriptive analysis for each of the six values of d, comparing the user

Descriptive Statistics								
Correlation difference (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.05	2.31	0.43	0.50	0.00	1.00	0.00	0.00	1.00
0.1	2.83	0.60	0.50	0.00	1.00	0.00	1.00	1.00
0.15	3.77	0.91	0.28	0.00	1.00	1.00	1.00	1.00
0.25	4.03	1.00	0.00	1.00	1.00	1.00	1.00	1.00
0.3	4.03	1.00	0.00	1.00	1.00	1.00	1.00	1.00
0.35	4.03	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Table 7.2: Descriptive Statistics for task JND Fine 0.3 showing Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Correlation difference (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.05	2.16	0.37	0.49	0.00	1.00	0.00	0.00	1.00
0.1	3.19	0.71	0.46	0.00	1.00	0.00	1.00	1.00
0.15	3.79	0.91	0.28	0.00	1.00	1.00	1.00	1.00
0.25	3.96	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.3	4.04	1.00	0.00	1.00	1.00	1.00	1.00	1.00
0.35	3.87	0.94	0.24	0.00	1.00	1.00	1.00	1.00

Table 7.3: Descriptive Statistics for task JND Fine -0.3 showing Mean Rank, Minimum, Maximum and Percentiles

performance based on the reference point chosen. The Friedman analysis reveals that there is no statistically significant difference between any RP while estimating higher values of d such as 0.35 ($\chi^2(3) = 2.400, p = 0.494$), 0.30 ($\chi^2(3) = 3.667, p = 0.300$) and 0.25 ($\chi^2(3) = 2.185, p = 0.536$). However, there is statistically significant difference between the four reference points while estimating lower values of d such as 0.05 ($\chi^2(3) = 11.676, p = 0.009$), 0.10 ($\chi^2(3) = 13.232, p = 0.004$) and 0.15 ($\chi^2(3) = 13.125, p = 0.004$).

7.3.3 Wilcoxon Test

We perform the Wilcoxon test to determine which pair of reference correlations are significantly different from each other when d has the value as 0.05, 0.1 and 0.15. The test reveals the following.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.05$ with reference correlation -0.3 and -0.7, ($Z = -3.162, p = 0.002$). Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference

Descriptive Statistics								
Correlation difference (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.05	1.83	0.20	0.41	0.00	1.00	0.00	0.00	0.00
0.1	3.11	0.63	0.49	0.00	1.00	0.00	1.00	1.00
0.15	3.71	0.83	0.38	0.00	1.00	1.00	1.00	1.00
0.25	4.14	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.3	4.06	0.94	0.24	0.00	1.00	1.00	1.00	1.00
0.35	4.14	0.97	0.17	0.00	1.00	1.00	1.00	1.00

Table 7.4: Descriptive Statistics for task JND Fine 0.7 showing Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Correlation difference (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.05	1.44	0.09	0.28	0.00	1.00	0.00	0.00	0.00
0.1	3.67	0.83	0.38	0.00	1.00	1.00	1.00	1.00
0.15	3.76	0.86	0.36	0.00	1.00	1.00	1.00	1.00
0.25	3.93	0.91	0.28	0.00	1.00	1.00	1.00	1.00
0.3	4.10	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.35	4.10	0.97	0.17	0.00	1.00	1.00	1.00	1.00

Table 7.5: Descriptive Statistics for task JND Fine -0.7 showing Mean Rank, Minimum, Maximum and Percentiles

of $d = 0.05$ with reference correlation 0.3 and -0.7, ($Z = -2.643, p = 0.008$). Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.10$ with reference correlation 0.3 and -0.7, ($Z = -3.130, p = 0.002$). Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.10$ with reference correlation 0.3 and -0.3, ($Z = -2.673, p = 0.008$). Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.15$ with reference correlation 0.3 and -0.7, ($Z = -2.324, p = 0.020$). Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.15$ with reference correlation 0.3 and 0.7, ($Z = -2.000, p = 0.046$).

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.16	0.09	0.28	0.00	1.00	0.00	0.00	0.00
0.7	2.39	0.20	0.41	0.00	1.00	0.00	0.00	0.00
-0.3	2.73	0.37	0.49	0.00	1.00	0.00	0.00	1.00
0.3	2.73	0.37	0.49	0.00	1.00	0.00	0.00	1.00

Table 7.6: Descriptive Statistics for task JND Fine with d=0.05 showing Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.86	0.83	0.38	0.00	1.00	1.00	1.00	1.00
0.7	2.46	0.63	0.49	0.00	1.00	0.00	1.00	1.00
-0.3	2.63	0.71	0.46	0.00	1.00	0.00	1.00	1.00
0.3	2.06	0.43	0.50	0.00	1.00	0.00	0.00	1.00

Table 7.7: Descriptive Statistics for task JND Fine with d=0.1 showing Mean Rank, Minimum, Maximum and Percentiles

Indeed the user performance is better when the latter is used as a reference point.

- There is a statistically significant difference while estimating a stimuli difference of $d = 0.15$ with reference correlation 0.3 and -0.3, ($Z = -2.840, p = 0.005$). Indeed the user performance is better when the latter is used as a reference point.

- For all the other remaining pairs, there is no significant difference between them.

7.3.4 Summary

The mean accuracy for all values of d at each reference point, RP, is summarized in Table 7.12. Analyzing the results therein, we reach the conclusion that the JND value for correlation estimation lies in the range 0.05 to 0.1 as most of the smaller mean values are concentrated in those two rows. Considering a 50% JND threshold level (which means the participant can reliably discriminate between two scatter plots 50% of the times), we obtain JND=0.05 as all the mean accuracy values are below 0.50. However, considering a tighter threshold of 75% JND level, results into a JND=0.10, as three out of four mean accuracy values are below 0.75.

We also observed that the user perception while estimating d is dependent on the reference correlation value. Consequently, we can conclude that the **JND value is**

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.61	0.86	0.36	0.00	1.00	1.00	1.00	1.00
0.7	2.56	0.83	0.38	0.00	1.00	1.00	1.00	1.00
-0.3	2.73	0.91	0.28	0.00	1.00	1.00	1.00	1.00
0.3	2.10	0.60	0.50	0.00	1.00	0.00	1.00	1.00

Table 7.8: Descriptive Statistics for task JND Fine with d=0.15 showing Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.44	0.91	0.28	0.00	1.00	1.00	1.00	1.00
0.7	2.56	0.97	0.17	0.00	1.00	1.00	1.00	1.00
-0.3	2.56	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.3	2.44	0.91	0.28	0.00	1.00	1.00	1.00	1.00

Table 7.9: Descriptive Statistics for task JND Fine with d=0.25 showing Mean Rank, Minimum, Maximum and Percentiles

dependent on the reference correlation value chosen. Having said that, we believe that the trend observed for JNDs may be participant dependent but considering the scope of this study, we postpone the in-depth study for future works.

7.4 Result Analysis for Weber's Task

The task for Weber observed the participant's accuracy while estimating correlation coefficient of scatter plots with an elliptical fit of data points. Figure 7.5 shows all the variations, for which the user accuracy was tested. We perform two kinds of analysis on the data, comparing participant performance for estimating positive correlations $(0.1, 0.2, 0.3, ..)$ and comparing participant estimates for positive correlations and their negative counter part (-0.9 vs. 0.9, -0.7 vs. 0.7..).

7.4.1 Absolute Error Analysis

Figure 7.6 shows the average error values for each of the positive correlation values (including $R = 0$) and Figure 7.7 depicts the graph comparing participant estimates for positive and negative odd correlation values.

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.49	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.7	2.43	0.94	0.24	0.00	1.00	1.00	1.00	1.00
-0.3	2.54	1.00	0.00	1.00	1.00	1.00	1.00	1.00
0.3	2.54	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Table 7.10: Descriptive Statistics for task JND Fine with d=0.3 showing Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Reference Correlation (d)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.7	2.50	0.97	0.17	0.00	1.00	1.00	1.00	1.00
0.7	2.50	0.97	0.17	0.00	1.00	1.00	1.00	1.00
-0.3	2.44	0.94	0.24	0.00	1.00	1.00	1.00	1.00
0.3	2.56	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Table 7.11: Descriptive Statistics for task JND Fine with d=0.35 showing Mean Rank, Minimum, Maximum and Percentiles

7.4.1.1 Graphical Analysis

As observed from the graph in Figure 7.6, the average error remains almost constant between $r = 0.2$ and $r = 0.6$ whereas the error rate is maximum on the extreme edges at higher correlation values. This is in contrast to the results of Bobko and Karren (1979) which stated that the error rate is most pronounced in the range $0.2 < |r| < 0.6$ compared to the boundary correlation values such as $r = \pm 0.9, \pm 0.8$. The large range bars for higher correlations such as $0.5, 0.6, 0.7, 0.8$ and 0.9 might give the false impression that the participants' estimates are spread across a wide range of values. However, in each one of them, the mean is much closer to 0.1 suggesting that the upper limit of the range, is due to the existence of few outliers. In fact, on

		Reference Correlation			
		-0.7	**-0.3**	**0.3**	**0.7**
Difference between two scatter plots (d)	**0.05**	0.09	0.37	0.43	0.20
	0.10	0.83	0.71	0.60	0.63
	0.15	0.86	0.91	0.91	0.83
	0.25	0.91	0.97	1.00	0.97
	0.30	0.97	1.00	1.00	0.94
	0.35	0.97	0.94	1.00	0.97

Table 7.12: Mean accuracy of all six scatter plot differences tested at four different reference correlation values

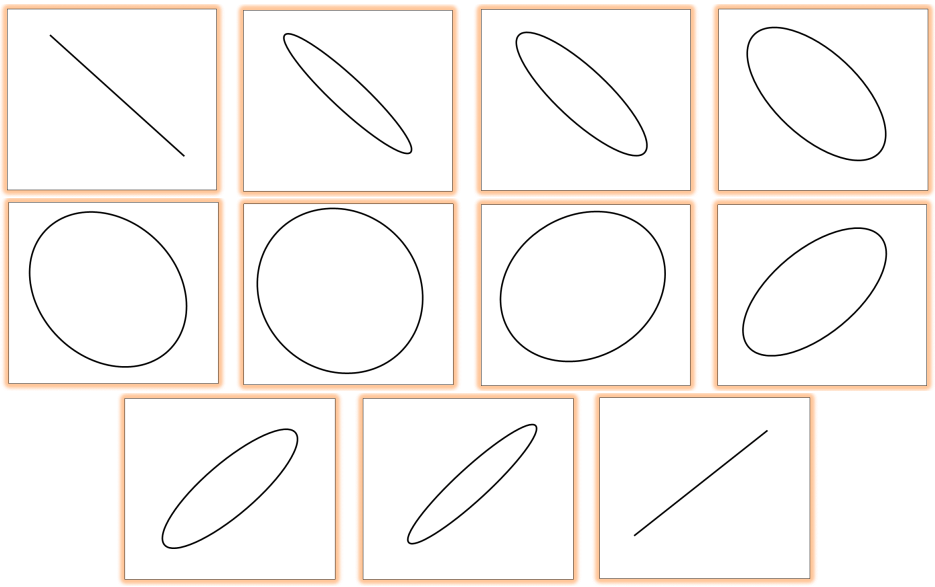


Figure 7.5: SCPs for Weber Task

closer examination of the data, it was found that only 1 or 2 (out of 35) participants contributed towards the upper limit in case of $r = 0.5, 0.6, 0.7, 0.8$ and 0.9.

The previously mentioned outlier effect is evident in the graph in Figure 7.7 also, for higher negative correlations with few participants making estimates, too far from the actual correlation value when $r = -0.5, -0.7, -0.9$. Further, upon comparing user estimates for positive and negative correlations, it becomes evident that there is no difference in average error when $r = \pm 0.1$ (average error = 0.09), $r = \pm 0.5$ (0.12) and $r = \pm 0.9$ (0.18). However, in case of $r = \pm 0.3, \pm 0.7$, for the same absolute correlation value, average error is more when direction of correlation is negative instead of positive. For example, when $r = -0.7$ average error is 0.17 whereas for $r = 0.7$ it is 0.14. We will perform the Friedman test to confirm whether these differences are significant or not.

From the two graphs, its easy to conclude that participant estimates are more precise for smaller correlation values, irrespective of their direction, as is evident from the small range in case of $r = \pm 0.1, \pm 0.3$. As the correlation value increases in either direction ($r = \pm 0.5, \pm 0.7, \pm 0.9$), the standard deviation gradually increases due to wide range of participant response.

Next we investigate the validity of Weber's law for estimating the correlation value from actual correlation value. If indeed Weber's law was valid, we would have $R_{est} = kR_{act}$, where R_{est} is the estimated correlation value, R_{act} is the actual correlation value and k is the Weber fraction and the graph would depict a straight line (see yellow dotted lines in Figure 7.8 for $k = 0.5, 1, 2$). However, Figure 7.8 clearly shows that the relation between estimated correlation and actual correlation (dotted red line) isn't linear. Thus, Weber's law isn't appropriate to model the perception of accuracy of

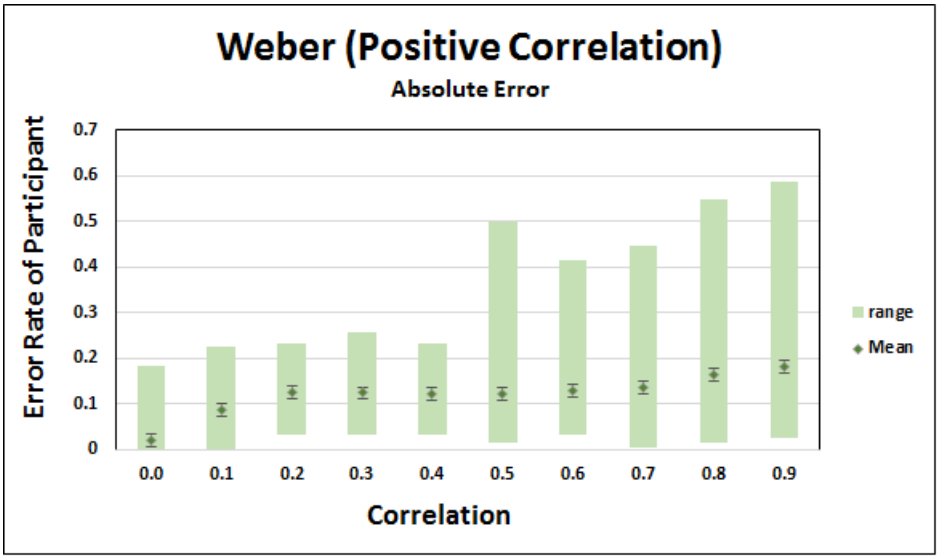


Figure 7.6: Performance analysis of positive correlations for task Weber

correlation estimation.

Similarly, Weber's law is inadequate to model the human perception of differences in correlation with respect to the objective differences in data correlation. If Weber's law was indeed applicable, then according to Harrison *et al.* (2014):

$$\Delta R_{est} = k\frac{\Delta R}{R}$$

where,

ΔR_{est} = change in perception or subjective estimates calculated as $R_{est}[i] - R_{est}[i-1]$, where i is any correlation

ΔR = objective difference in data correlation

R = overall correlation in data

k = Weber's fraction

We consider a fixed step size of 0.1 i.e. each time the objective correlation increments by 0.1 starting from 0 to 0.9. Also, k being the Weber's constant has a fixed value. Thus, we obtain the relation:

$$\Delta R_{est} \propto \frac{1}{R}$$

Thus, there is an inverse relation between change in perception (subjective correlation) and the base correlation. This equation is equivalent to $y = \frac{1}{x}$ and hence, ideally, for different values of the Weber fraction k, the graph should have a horizontal asymptote at $y = 0$ and vertical asymptote at $x = 0$ (see Figure 7.9 (a)). Instead plotting the graph from participant data (see Figure 7.9(b)), we obtain a rather random curve that doesn't fit the ideal Weber's curve. These results are in contradiction to those of Harrison *et al.* (2014) which stated that the Weber's law is applicable.

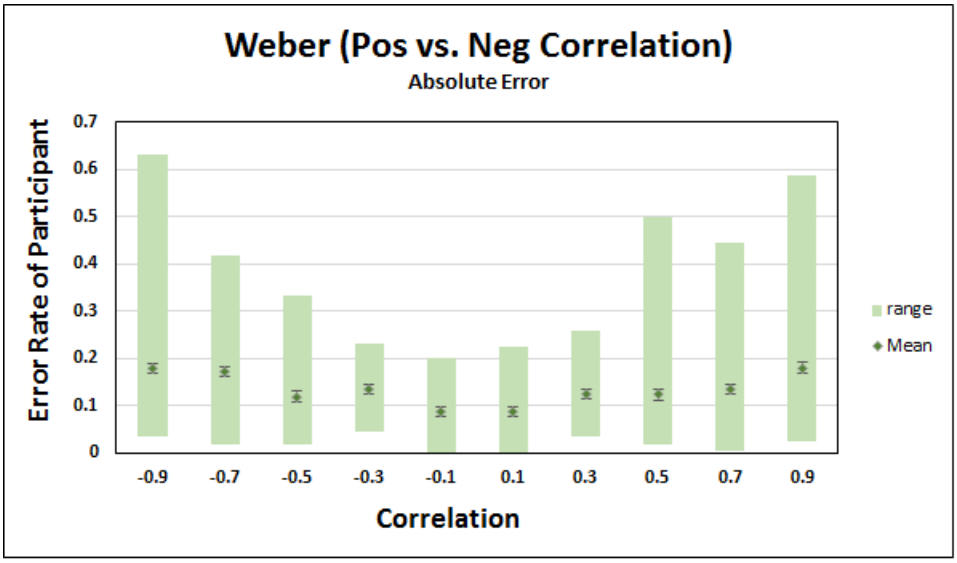


Figure 7.7: Comparison of user performance for positive vs. negative correlation in task Weber

7.4.1.2 Friedman Analysis

We further support our observations deduced from the graph using the Friedman test analysis, which suggests there is a statistically significant difference in average error while estimating positive correlations excluding $R \approx 0$, $(\chi^2(8) = 26.359, p = 0.000)$. We exclude R because its graphically clear that it has the lowest error rate compared to all other values and hence would be significant with respect to all other correlation values. The descriptive analysis of the positive correlations in given in Table 7.13. The Friedman mean-ranking (see Table 7.13) further gives us an estimate of the ordering of positive correlation values based on the accuracy of user estimation, with the maximum value of average error for $R = 0.9$. Table 7.14 represents the descriptive statistics for positive vs. negative correlation values obtained from the Friedman analysis which reveals that an overall significance is present $(\chi^2(9) = 52.661, p = 0.000)$.

7.4.1.3 Wilcoxon Test

Further, we perform post-hoc test of *Wilcoxon signed-rank* analysis, to check which pair of positive and negative correlation $(+R, -R)$ are in particular significantly different from each other. We perform pairwise comparison of all the negative correlations with their positive counterpart. The Wilcoxon signed-rank test reports the following:

- There is no statistically significant difference in user estimation of correlation for 0.9 or -0.9, $(Z = -1.557, p = 0.120)$. Indeed, median error rate at both the levels was 0.14 and 0.15 respectively, which are not very different from each other.

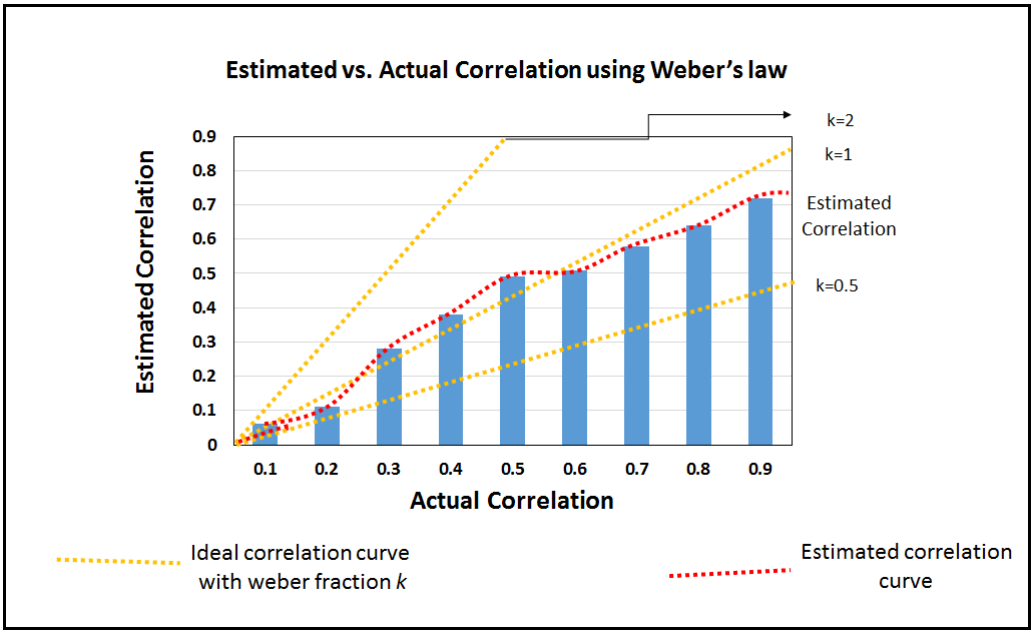
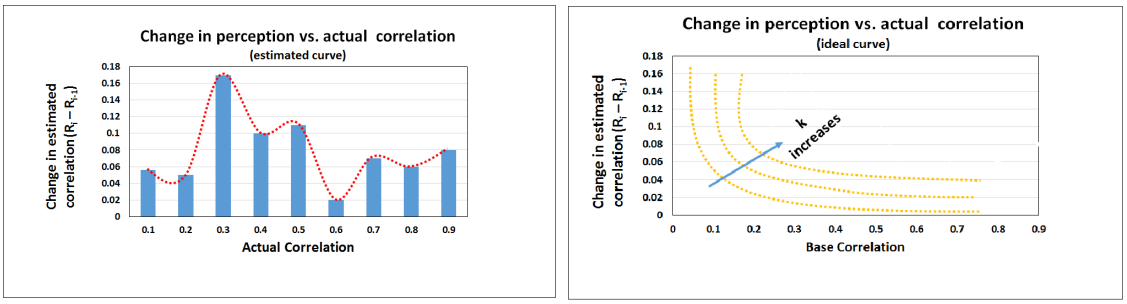


Figure 7.8: Proving invalidity of Weber's law for determining subjective correlation from objective correlation values



(a) Observed curve for human perception of correlation difference

(b) Ideal Weber's curve for human perception of correlation difference for various values of k

Figure 7.9: Graphs comparing human perception of differences in correlation with actual differences in correlation.

- There is a statistically significant difference in user estimation of correlation for 0.7 and -0.7, ($Z = -3.104, p = 0.002$). However, the performance of the former is slightly better compared to the latter.

- There is no statistically significant difference in user estimation of correlation for 0.5 or -0.5, ($Z = -0.371, p = 0.711$). Indeed, median error rate at both the levels was 0.67.

- There is no statistically significant difference in user estimation of correlation for 0.3 or -0.3, ($Z = -1.813, p = 0.070$). Indeed, median error rate at both the levels was 0.12 and 0.14 respectively, which are not very different from each other. This means that the difference in average error obtained during graphical analysis in Section 7.4.1.1 was only by chance.

Descriptive Statistics								
Correlation (r)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.1	3.39	0.09	0.05	0.00	0.23	0.05	0.09	0.10
0.2	5.30	0.13	0.06	0.03	0.23	0.07	0.13	0.17
0.3	5.16	0.12	0.06	0.03	0.26	0.09	0.12	0.15
0.4	4.90	0.12	0.05	0.03	0.23	0.07	0.12	0.17
0.5	4.49	0.12	0.09	0.02	0.50	0.07	0.10	0.15
0.6	4.70	0.13	0.08	0.03	0.42	0.07	0.10	0.18
0.7	4.90	0.14	0.09	0.00	0.45	0.06	0.12	0.20
0.8	5.81	0.16	0.11	0.02	0.55	0.08	0.15	0.23
0.9	6.36	0.18	0.14	0.03	0.59	0.09	0.14	0.23

Table 7.13: Descriptive Statistics for task Weber comparing positive correlations showing Mean Rank, Maximum, Minimum and Percentiles

Descriptive Statistics								
Correlation (r)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
-0.9	7.03	0.18	0.11	0.03	0.63	0.12	0.15	0.25
-0.7	7.00	0.17	0.10	0.02	0.42	0.10	0.16	0.25
-0.5	5.36	0.12	0.07	0.02	0.33	0.07	0.12	0.15
-0.3	6.21	0.14	0.05	0.04	0.23	0.10	0.14	0.17
-0.1	3.54	0.09	0.04	0.00	0.20	0.05	0.10	0.10
0.1	3.66	0.09	0.05	0.00	0.23	0.05	0.09	0.10
0.3	5.33	0.12	0.06	0.03	0.26	0.09	0.12	0.15
0.5	5.04	0.12	0.09	0.02	0.50	0.07	0.10	0.15
0.7	5.30	0.14	0.09	0.00	0.45	0.06	0.12	0.20
0.9	6.53	0.18	0.14	0.03	0.59	0.09	0.14	0.23

Table 7.14: Descriptive Statistics for task Weber comparing positive and negative correlations and showing Mean Rank, Maximum, Minimum and Percentiles

- There is no statistically significant difference in user estimation of correlation for 0.1 or -0.1, ($Z = -0.013, p = 0.989$). Indeed, median error rate at both the levels was 0.09 and 0.10 respectively, which are not very different from each other.

In addition, we also compare user performance at all positive correlations to find pair of correlation values, significantly different from each other. Table 7.15 contains the relevant test statistics for positive correlation value. We observe that there is a significant difference in median error rate when one of the scatter plot is least positively correlated ($R = 0.1$) compared to the other and user performance while estimating correlation is better for the former. We expect to see such a trend because when $R = 0.1$, the scatter plot has a random distribution of data points in the scatter plots

Test Statistics													
	0.2 - 0.1	0.3 - 0.1	0.4 - 0.1	0.6 - 0.1	0.7 - 0.1	0.8 - 0.1	0.9 - 0.1	0.8 - 0.4	0.9 - 0.5	0.8 - 0.6	0.9 - 0.6	0.8 - 0.7	0.9 - 0.7
Z	-3.10	-2.83	-2.49	-2.32	2.58	-3.56	-4.04	-1.97	-2.48	-2.18	-2.48	-2.06	-2.62
Asymp Sig. (2-tailed)	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.05	0.01	0.03	0.01	0.04	0.01

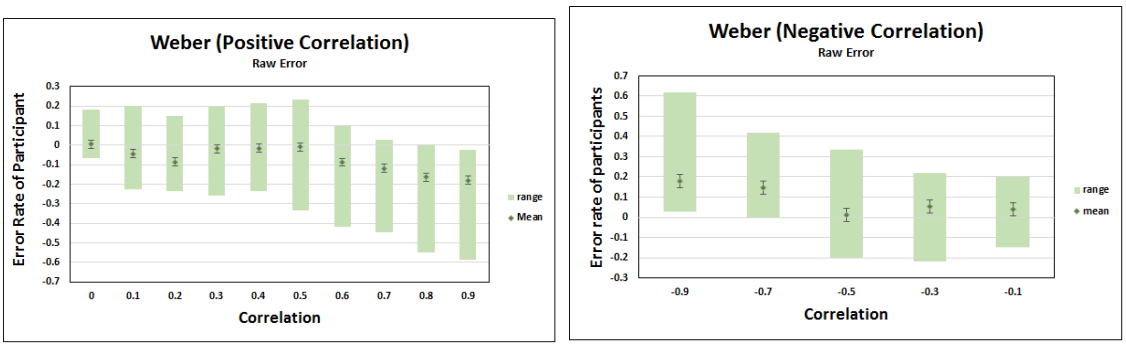
Table 7.15: Test Statistics for user estimation of positive correlation values.

which is not difficult to detect. Similarly, since the very strongly positive correlation ($R = 0.9$) are easily distinguishable compared to other moderately high correlation $(0.5, 0.6, 0.7)$, there is a significantly different in their error rates. Lastly, we see a intriguing pattern which suggests that the correlation estimation is better when $R = 0.8$ compared to $R = 0.7$, although they both differ by an amount of 0.1 only. This suggests that human perception improves as the pattern of data points becomes more prominent at higher correlation values.

The rest of the pairs have no statistically significant difference between them, and are therefore excluded from mentioning here.

7.4.2 Raw Error Analysis

Figure 7.10 shows the graph plotted with raw error values. We observe an overestimation for all the negative correlation values (in addition to 0). On the contrary, an underestimation is observed for all the positive correlation values. There seems to be a irregular pattern of error rate in both the directions of correlation and we obtain a non linear curve in the graph. The error rate is the highest at the two ends i.e. $r = -0.9$ and $r = 0.9$ where the correlation value is the highest. The participant performance seems to be exceptionally accurate at $r = \pm 0.5$ which is odd.



(a) Performance analysis of positive correlations (b) Performance analysis of negative correlations

Figure 7.10: Performance analysis of task Weber using raw error values

7.4.3 Summary

Thus, we can conclude that **human perception** is affected by a variation in correlation coefficient, even in case of regular elliptical cloud shape, and hence is **unreliable when estimating correlation** from scatter plots. Also, the use Weber's model isn't appropriate in context of estimating correlation for both positive and negative values as we observed the estimated curve is far from the ideal Weber's curve. Lastly for similar absolute values of correlation but different directions, the user performance is consistent except in case of $r = \pm 0.7$.

7.5 Result Analysis for Distribution Task

The task for Distribution observed the effect of varying the cloud shape by distributing all the data points in three clusters: two circles along the -45 degree line and an ellipse along the 45 degree line. Figure 7.12 shows all the variations, for which the user accuracy was tested. We use, as the reference point, the independent level which corresponds to the scatter plot that has all the data points inside 1 elliptical cluster only i.e. absence of the two circles on the -45 degree line and thus a distribution ratio of $0 : 80 : 0$ (see highlighted scatter plot in Figure 7.12), against which we compare the variations in distribution level of the scatter plots.

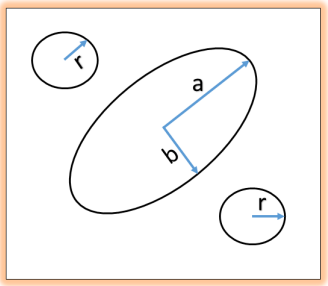


Figure 7.11: Labelled scatter plot for task Distribution

7.5.1 Absolute Error Analysis

Figure 7.13 shows the average accuracy for each level of the independent variable, which in this case is the distribution ratio that divides all the data points amongst the two circles and the ellipse, in different proportions (see Figure 7.11).

7.5.1.1 Graphical Analysis

As explained in Section 4.2.3, each unique distribution ratio, corresponds to a different correlation value. As a result, the average error for each successive level of distribution

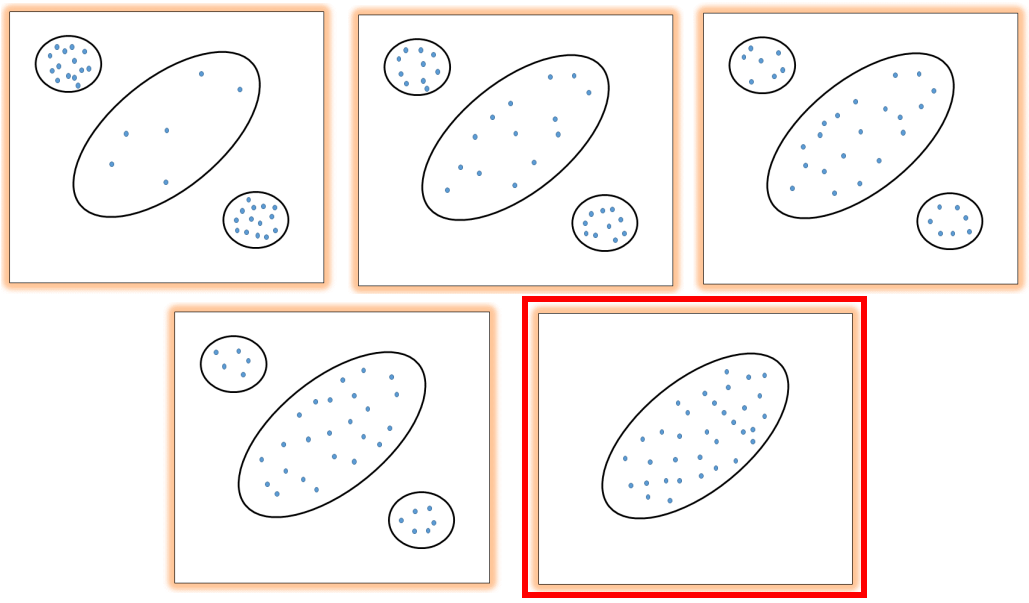


Figure 7.12: Scatter plots for task Distribution

ratio might be influenced not only by change in distribution ratio but also by change in correlation coefficient. Thus, it becomes necessary to reduce the confounding effect due to the change in correlation coefficient while analyzing the user performance for this task by isolating the effects of correlation and distribution ratio. Hence, while plotting the average error for each distribution ratio level (in green dots), we plot along with each of these, the average error for the correlation value, (corresponding to the distribution level), estimated by the participants when there aren't any clusters on the -45 degree line i.e. elliptical point cloud (see the red '×' symbols), as illustrated in Figure 7.13. As previously observed in Section 6.4, the human perception is unreliable and the average accuracy varies depending on change in correlation coefficient level. However if there was no effect of the distribution ratio on the human perception of correlation estimation, the two error rates for each correlation value would be identical. But this is not the case, as we observe a vast difference in green and red mean values for a particular correlation coefficient, thereby validating that the change in error rate (in green) is majorly due to change in distribution ratio levels.

The error rate, as is observed from the chart, gradually increases with a corresponding increase in the level of the distribution ratio from the reference case (where distribution ratio is $0 : 80 : 0$). The graph depicts a sudden increase in average error from 0.16 at $5 : 70 : 5$ to 0.46 at $10 : 60 : 10$ indicating a significant difference between these two distribution ratio levels. The minimum, maximum value and standard error for each level is depicted in the Figure 7.13.

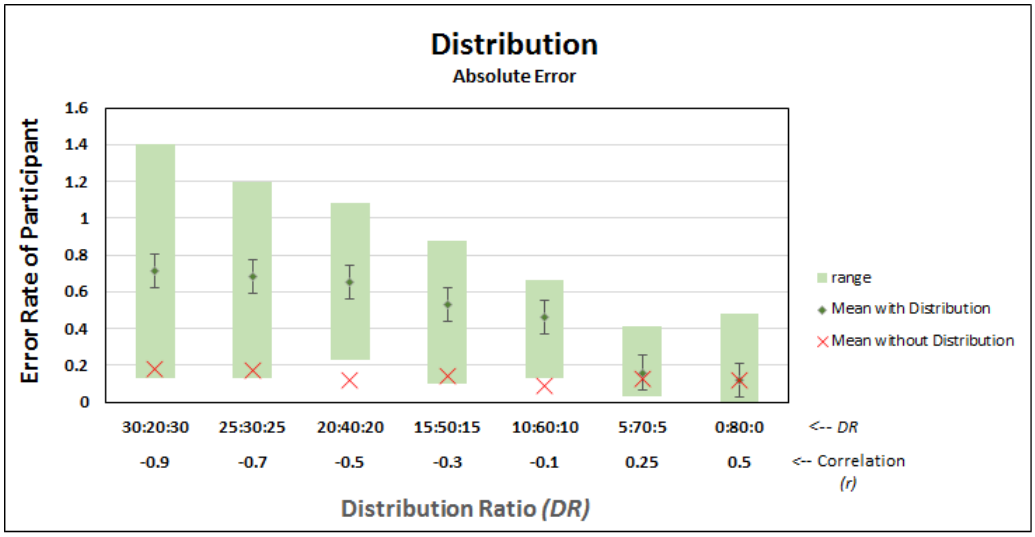


Figure 7.13: Performance Analysis for task Distribution with absolute error. Since, we did not collect the user estimate at $R = 0.25$ for a scatter plot without any distribution, we use the average of user estimate at $R = 0.2$ and $R = 0.3$

7.5.1.2 Friedman Analysis

We further support our observations deduced from the graph using *Friedman test* analysis, which suggests there is a statistically significant difference in average error values while estimating correlation, depending on the levels of distribution ratio, $(\chi^2(6) = 133.350, p = 0.000)$. The descriptive analysis of this task in given in Table 7.16. The Friedman mean-ranking (see Table 7.16) further gives us an estimate of the ordering of independent levels based on the **accuracy** of user estimation with the maximum average error corresponding to the distribution ratio $25 : 30 : 25$.

Descriptive Statistics								
Distribution Ratio	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
30:20:30	5.57	0.71	0.32	0.13	1.40	0.37	0.73	0.93
25:30:25	5.60	0.68	0.30	0.13	1.20	0.40	0.70	0.93
20:40:20	5.34	0.65	0.23	0.23	1.08	0.48	0.60	0.85
15:50:15	4.30	0.53	0.18	0.10	0.88	0.42	0.50	0.68
10:60:10	3.80	0.46	0.13	0.13	0.67	0.33	0.50	0.58
5:70:5	1.76	0.16	0.09	0.03	0.42	0.08	0.15	0.22
0:80:0	1.63	0.12	0.09	0.02	0.50	0.07	0.10	0.15

Table 7.16: Descriptive Statistics for task Distribution with Mean Rank, Minimum, Maximum and Percentiles

7.5.1.3 Wilcoxon Test

Further, we perform post-hoc test of *Wilcoxon signed-rank* analysis, to check which groups are in particular significantly different from each other. We perform pair-wise comparison of all independent levels with the reference independence level. The Wilcoxon signed-rank test shows the following results:

- There is no statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 5 : 70 : 5, ($Z = -1.830, p = 0.067$). Indeed, median error rate at both the levels was 0.15 and 0.10 respectively, which are not very different from each other.
- There is a statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 10 : 60 : 10, ($Z = -5.087, p = 0.000$). However, the performance of former is better compared to latter.
- There is a statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 15 : 50 : 15, ($Z = -5.111, p = 0.000$). However, the performance of former is better compared to latter.
- There is a statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 20 : 40 : 20, ($Z = -5.070, p = 0.000$). However, the performance of former is better compared to latter.
- There is a statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 25 : 30 : 25, ($Z = -5.070, p = 0.000$). However, the performance of former is better compared to latter.
- There is a statistically significant difference between a distribution ratio of 0 : 80 : 0 compared to 30 : 20 : 30, ($Z = -5.036, p = 0.000$). However, the performance of former is better compared to latter.

7.5.2 Raw Error Analysis

Figure 7.14 shows the graph plotted with raw error values. We observe an increasing pattern with increase in distribution level similar to the graph plotted with absolute error. We observe that on an average, participants generally tend to overestimate correlation values for all the distribution ratio levels. However, this overestimation is the highest when the distribution of points in clusters along the -45 degree line is highest i.e. a distribution ratio equal to 30 : 20 : 30 or 25 : 30 : 20, as compared to the average error at other levels.

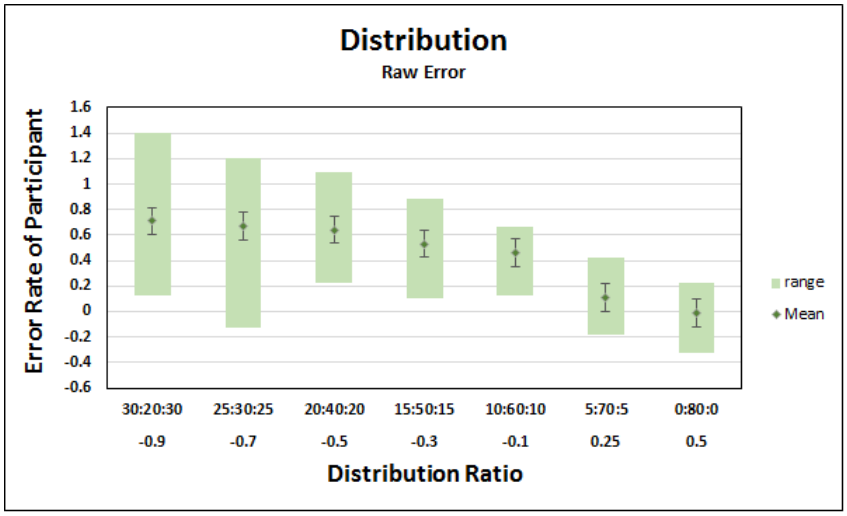


Figure 7.14: Performance analysis for task Distribution with raw error

7.5.3 Summary

Thus, we can conclude that **human perception is affected by a variation in distribution of points in the scatter plot** along the 45 degree and -45 degree line. We can also conclude that the correlation coefficient statistic itself, as a standalone indicator of correlation in a scatter plot, is unreliable. As observed from this task, a single correlation value can corresponds to (at least) two different scatter plots for which the user accuracy is always different.

7.6 Result Analysis for Density Task

The task for Density observed the effect of varying the number of data points used to plot a scatter plot. We measure five different density variations at three different correlation levels. Figure 7.15 gives examples of all the density variations at which the user accuracy was tested.

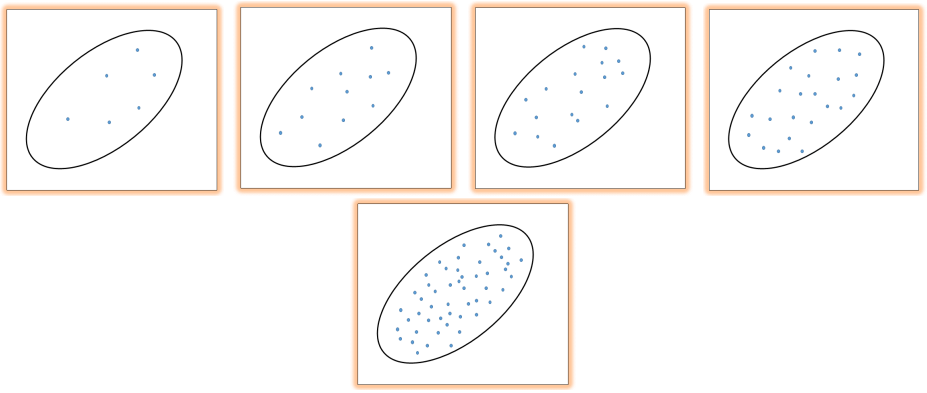


Figure 7.15: Scatter plots for task Density

7.6.1 Absolute Error Analysis

Figure 7.16 shows the average error rate for each level of density at all the three reference correlation values.

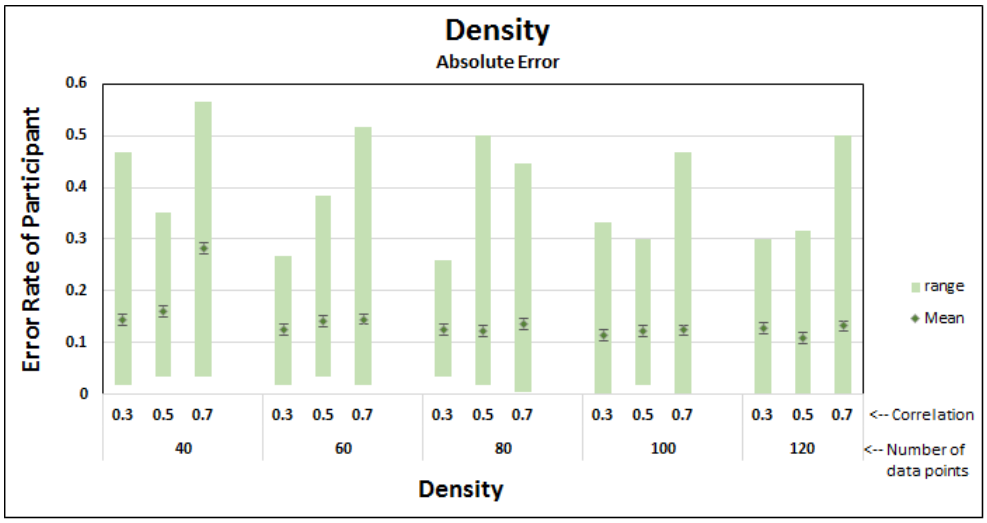


Figure 7.16: Performance analysis of task Density using absolute error

7.6.1.1 Graphical Analysis

The graph depicts that the error rate is maximum(0.28) when 40 data points are used to plot a scatter plot of correlation, $R = 0.7$. For the remaining correlation values the means of the error at different density levels is constant.

7.6.1.2 Friedman Analysis

We support our observations, deduced from the graph using Friedman test analysis, which suggests that there is a statistically significant difference between the different density level, only for the correlation value, $R = 0.7$ ($\chi^2(4) = 63.670, p = 0.000$). For the remaining correlation values, the Friedman analysis reports no significant difference at any correlation level. The descriptive analysis for each correlation group is given in Table 7.17, Table 7.18 and Table 7.19. In case of $R = 0.7$, the Friedman mean-ranking further suggests that the maximum participant error is observed when 40 data points are used to plot a scatter plot of correlation value 0.7. Also, the most optimum number of data points required to plot the same correlation is 100. Table 7.17 and Table 7.18 suggest that to plot $R = 0.3$ and $R = 0.5$, the optimum number of data points are 100 and 120 respectively.

Descriptive Statistics								
Density (no. of data points)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
40	3.41	0.14	0.09	0.02	0.47	0.07	0.12	0.17
60	2.93	0.12	0.07	0.02	0.27	0.07	0.10	0.20
80	2.97	0.12	0.06	0.03	0.26	0.09	0.12	0.15
100	2.74	0.11	0.07	0.00	0.33	0.07	0.10	0.15
120	2.94	0.13	0.07	0.00	0.30	0.07	0.13	0.15

Table 7.17: Descriptive Statistics for task Density at R=0.3 with Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
Density (no. of data points)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
40	3.51	0.16	0.09	0.03	0.35	0.10	0.12	0.23
60	3.19	0.14	0.08	0.03	0.38	0.08	0.12	0.18
80	2.89	0.12	0.09	0.02	0.50	0.07	0.10	0.15
100	2.87	0.12	0.07	0.02	0.30	0.07	0.12	0.17
120	2.54	0.11	0.06	0.00	0.32	0.07	0.10	0.13

Table 7.18: Descriptive Statistics for task Density at R=0.5

7.6.1.3 Wilcoxon Test

The Wilcoxon test is performed to determine which groups of density levels are statistically different from each other while estimating a correlation coefficient equal to 0.7. We perform pairwise comparisons of all levels to get the following result.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 80 or 60 points ($Z = -1.034, p = 0.301$). Indeed they have almost similar median error rate of 0.12 and 0.13, respectively, as can be seen in Table 7.19.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 100 or 60 points ($Z = -1.744, p = 0.081$). Indeed they have almost similar median error rate of 0.08 and 0.13, respectively, as can be seen in Table 7.19.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 120 or 60 points ($Z = -0.787, p = 0.431$). Indeed they have equal median error rate of 0.13, as can be seen in Table 7.19.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 100 or 80 points ($Z = -1.442, p = 0.149$). Indeed

Descriptive Statistics								
Density (no. of data points)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
40	4.84	0.28	0.14	0.03	0.57	0.18	0.27	0.35
60	2.86	0.15	0.11	0.02	0.52	0.07	0.13	0.18
80	2.56	0.14	0.09	0.00	0.45	0.06	0.12	0.20
100	2.17	0.12	0.10	0.00	0.47	0.05	0.08	0.18
120	2.57	0.13	0.11	0.00	0.50	0.05	0.13	0.20

Table 7.19: Descriptive Statistics for task Density at R=0.7

they have almost similar median error rate of 0.08 and 0.12, respectively, as can be seen in Table 7.19.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 120 or 80 points ($Z = -0.696, p = 0.486$). Indeed they have almost similar median error rate of 0.13 and 0.12, respectively, as can be seen in Table 7.19.

- There is no statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 120 or 100 points ($Z = -0.579, p = 0.563$). Indeed they have almost similar median error rate of 0.13 and 0.08, respectively, as can be seen in Table 7.19.

- There is a statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 60 and 40 points ($Z = -5.021, p = 0.000$). Moreover, user performance is better at former density level.

- There is a statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 80 and 40 points ($Z = -4.947, p = 0.000$). Moreover, user performance is better at former density level.

- There is a statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 100 and 40 points ($Z = -5.113, p = 0.000$). Moreover, user performance is better at former density level.

- There is a statistically significant difference between plotting a scatter plot of correlation coefficient 0.7 with 120 and 40 points ($Z = -5.031, p = 0.000$). Moreover, user performance is better at former density level.

7.6.2 Raw Error Analysis

Figure 7.17 shows the graph with raw error values of the participant estimation of correlation. The means as well as the range for most of the levels lie below the

positive y-axis, irrespective of the correlation value, which clearly shows that most of the participants tend to underestimate correlation in scatter plots plotted with varying number of data points, with only a few minor exceptions of overestimation. The average error of underestimation is highest when plotting scatter plot with 40 data points and $r = 0.7$. Lastly, the graph clearly shows that within a particular density level, users tend to underestimate more at higher correlation values as is depicted by the greater mean error values.

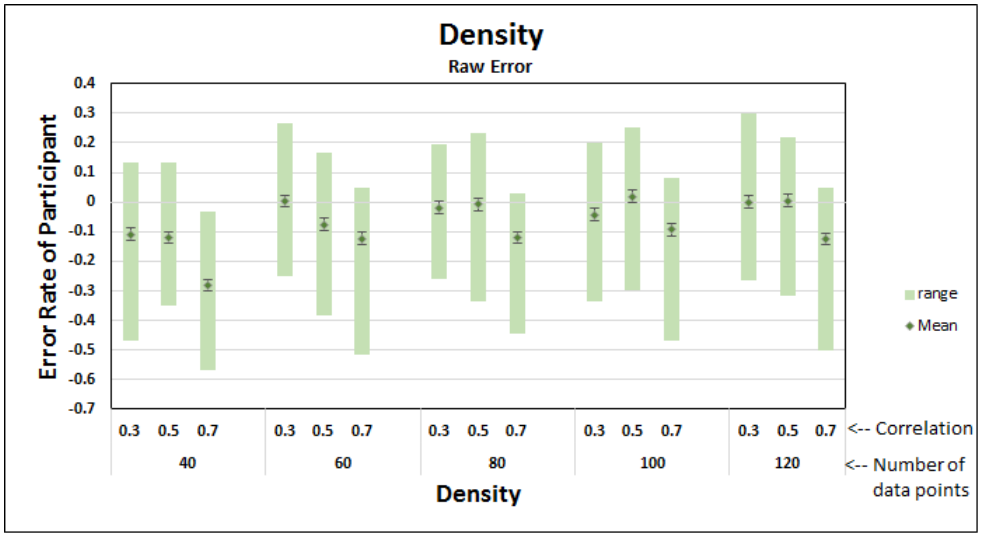


Figure 7.17: Performance analysis for task Density with raw error values

7.6.3 Summary

Thus, we conclude that **human perception is affected by a variation in density of the scatter plot** i.e. number of data points used to plot a scatter plot and the effect is more pronounced when a scatter plot with high correlation coefficient is plotted with few data points.

7.7 Result Analysis for Reflective Asymmetry Task

The task for Reflective Asymmetry observed the effect of varying the point cloud shape in scatter plots such that data points lie on either side of the 45 degree line but are reflectively asymmetrical as described in Figure 7.18 which shows all the variations for which user accuracy is tested at a particular correlation value. We use, as the reference point, the independence level which corresponds to the scatter plot that displays uniform symmetry on either side of the 45 degree line (see highlighted scatter plots in Figure 7.19), against which we compare the variations in reflective asymmetry of the scatter plots displaying a particular correlation value.

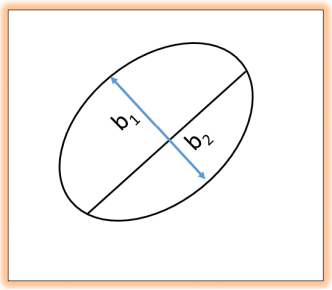


Figure 7.18: Labelled scatter plot for task Reflective Asymmetry

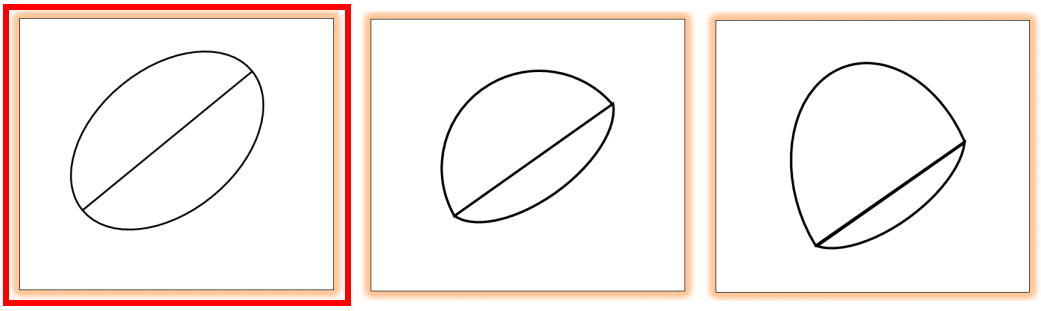


Figure 7.19: Different levels of the task Reflective asymmetry analyzed with reference level scatter plot with in red border

7.7.1 Absolute Error Analysis

Figure 7.20 shows the average error value of participants for each level of the independent variables, which in this case is the pair of semi-minor axes (b_1, b_2) of the two half ellipses, at three correlation values 0.5, 0.7 and 0.9 (see Figure 7.19)

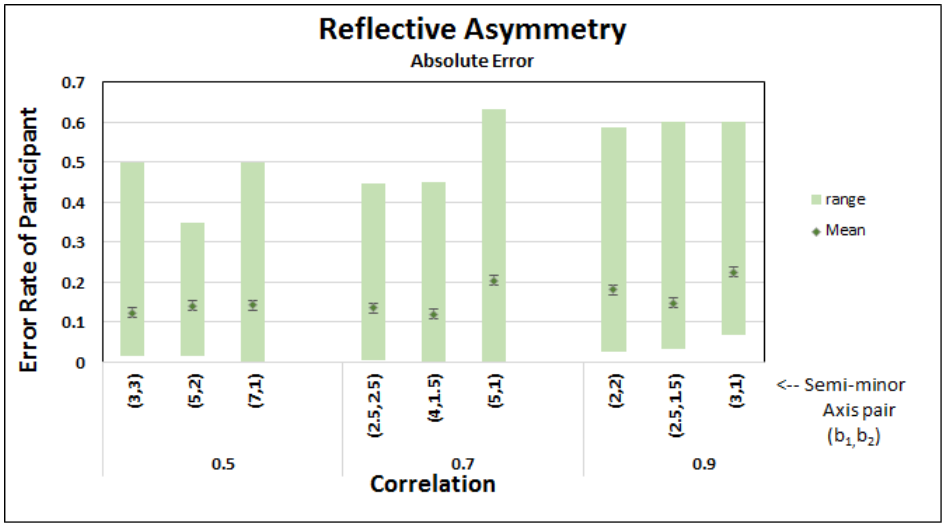


Figure 7.20: Performance analysis for task Reflective asymmetry with absolute error

7.7.1.1 Graphical Analysis

As observed from the graph, the mean error value for all reflective asymmetry levels are same when $R = 0.5$. This means a variation in reflective asymmetry level for lower correlation doesn't have an impact on user perception of correlation estimation. However, we see a significant difference in mean error values as the correlation increases to $R = 0.5$ and further to $R = 0.9$. The minimum, maximum value and standard error for each is depicted in the Figure 7.20.

7.7.1.2 Friedman Analysis

We verify our observations, derived from the graph, with the help of Friedman analysis which suggest there is a statistical difference in average error in correlation estimation at higher correlations values of $R = 0.7$, $(\chi^2(2) = 13.956, p = 0.001)$ and $R = 0.9$, $(\chi^2(2) = 24.827, p = 0.000)$. As is evident from the p-value of the two Friedman analysis, the significant difference is more for the latter correlation value as compared to the former. The descriptive analysis of this task for each correlation value is given in Table 7.20, 7.21 and 7.22, respectively. The Friedman ranking suggests that within the higher correlation group ($R = 0.7$ and $R = 0.9$), the error rate is highest when the scatter plot is highly reflective asymmetrical.

Descriptive Statistics								
(b_1, b_2)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
(3, 3)	1.91	0.10	0.08	0.00	0.33	0.05	0.07	0.15
(5, 2)	2.09	0.13	0.10	0.00	0.35	0.03	0.12	0.20
(7, 1)	2.00	0.13	0.12	0.00	0.50	0.03	0.08	0.18

Table 7.20: Descriptive Statistics for task Reflective Asymmetry at $R = 0.5$ with Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
(b_1, b_2)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
(2.5, 2.5)	1.86	0.12	0.10	0.00	0.45	0.03	0.12	0.20
(4, 1.5)	1.63	0.11	0.11	0.00	0.45	0.02	0.08	0.20
(5, 1)	2.51	0.20	0.16	0.00	0.63	0.07	0.15	0.33

Table 7.21: Descriptive Statistics for task Reflective Asymmetry at $R = 0.7$ with Mean Rank, Minimum, Maximum and Percentiles

Descriptive Statistics								
(b_1, b_2)	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
(2, 2)	1.84	0.18	0.14	0.03	0.59	0.09	0.14	0.23
(2.5, 1.5)	1.51	0.15	0.11	0.03	0.60	0.08	0.12	0.18
(3, 1)	2.64	0.23	0.13	0.07	0.60	0.12	0.20	0.33

Table 7.22: Descriptive Statistics for task Reflective Asymmetry at $R = 0.9$ with Mean Rank, Minimum, Maximum and Percentiles

7.7.1.3 Wilcoxon Test

To detect the source of the significance, we perform next, the *Wilcoxon test* to detect which level of reflective asymmetry generated the maximum error at a particular correlation value. Thus, we pairwise analyze all the three reflective symmetry levels for both the correlation values, $R = 0.7$ and $R = 0.9$. It reveals the following results:

- For $R = 0.7$, there is statistically significant difference between optimum value at $(b_1 = 2.5, b_2 = 2.5)$ and independent level of $(b_1 = 5, b_2 = 1)$, $(Z = -3.260, p = 0.001)$. Out of the two, user performance is better for the former independent level.

- For $R = 0.7$, there is no statistically significant difference between optimum value at $(b_1 = 2.5, b_2 = 2.5)$ and independent level of $(b_1 = 4, b_2 = 1.5)$, $(Z = -1.286, p = 0.198)$.

- For $R = 0.7$, there is statistically significant difference between independent levels $(b_1 = 5, b_2 = 1)$ and $(b_1 = 4, b_2 = 1.5)$, $(Z = -3.765, p = 0.000)$. Out of the two, user performance is better for the latter independent level.

- For $R = 0.9$, there is statistically significant difference between optimum value at $(b_1 = 2, b_2 = 2)$ and independent level of $(b_1 = 3, b_2 = 1)$, $(Z = -3.686, p = 0.000)$. Out of the two, user performance is better for the former independent level.

- For $R = 0.9$, there is no statistically significant difference between optimum value at $(b_1 = 2, b_2 = 2)$ and independent level of $(b_1 = 2.5, b_2 = 1.5)$, $(Z = -1.779, p = 0.075)$.

- For $R = 0.9$, there is statistically significant difference between independent levels $(b_1 = 3, b_2 = 1)$ and $(b_1 = 2.5, b_2 = 1.5)$, $(Z = -4.649, p = 0.000)$. Out of the two, user performance is better for the latter independent level.

7.7.2 Raw Error Analysis

Figure 7.21 shows the graph plotted with average error calculated using raw errors of the participants. For each correlation value, a significantly large portion of range and the subsequent mean error value of all reflective asymmetry levels, lying in the negative y-axis, suggests that users tend to underestimate correlation values in general, irrespective of the reference correlation value. Also, within each group, the underestimation is maximum when the level of reflective asymmetry is the highest (compared to the reference independent level). Lastly, we observe that the mean user underestimation while estimating correlation, keeps on increasing gradually as the reference correlation increases.

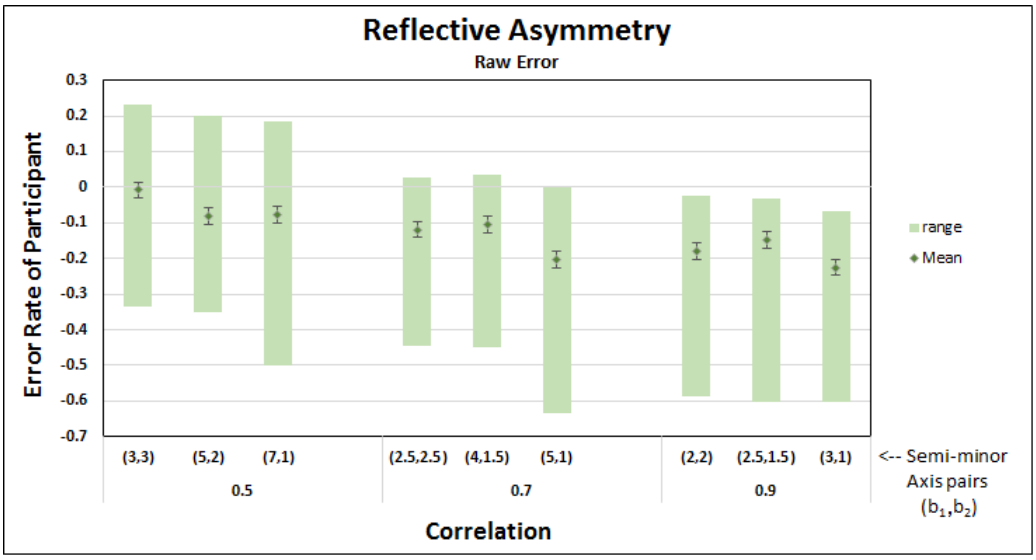


Figure 7.21: Performance analysis of task Reflective asymmetry with raw error

7.7.3 Summary

We conclude that **reflective asymmetry tends to affect user perception** for higher correlation values and as a result, the users tend to underestimate most correlation values. The tendency to make an error is most pronounced when the level of the reflective asymmetry is the highest.

7.8 Result Analysis for Progressive Symmetry Task

The task for Progressive Symmetry observed the effect of varying the point cloud shape in scatter plot in such a way that a cluster of points (in form of a circle) is placed at the top of ellipse (placed along the 45 degree line). Figure 7.23 which shows all the variations for which the user accuracy was tested. We use, as the reference

point, the independent level which corresponds to the scatter plot with elliptical fit of data i.e. absence of circle at the top (see highlighted scatter plot in Figure 7.23), against which we compare the variations in progressive symmetry of the scatter plot.

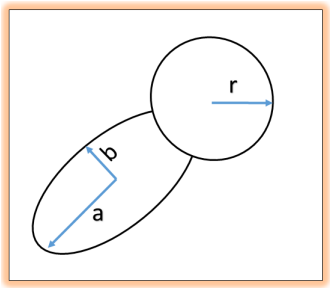


Figure 7.22: Labelled scatter plot for task Progressive Symmetry

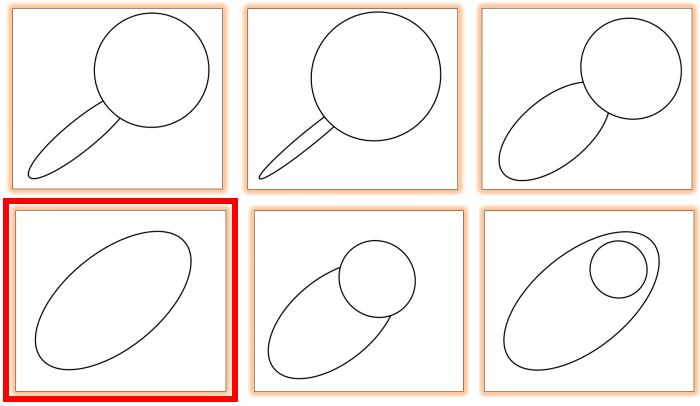


Figure 7.23: Different levels of task Progressive symmetry with reference level scatter plot in red border

7.8.1 Absolute Error Analysis

Figure 7.24 shows the average absolute error for each independent level of the progressive symmetry, which in this case is the semi-minor axis of the ellipse (b) (see Figure E.2).

7.8.1.1 Graphical Analysis

The accuracy, as is observed from the chart, is highest in the middle at $b = 2.5$ with lowest amount of average error (0.14), which is the reference case. The graph shows a curvilinear U-shaped curve indicating highest error rate (0.32 at $b = 0.5$ and 0.36 at $b = 3.5$) at the two extreme ends. This is indicative of the fact that as the progressive symmetry of a symmetric scatter plot ($b = 2.5$) varies in either direction (towards $b = 3.5$ or $b = 0.5$), the user perception is affected and the average accuracy goes down as average error rate goes up. The minimum, maximum value and standard error for each level is depicted in the Figure 7.25.

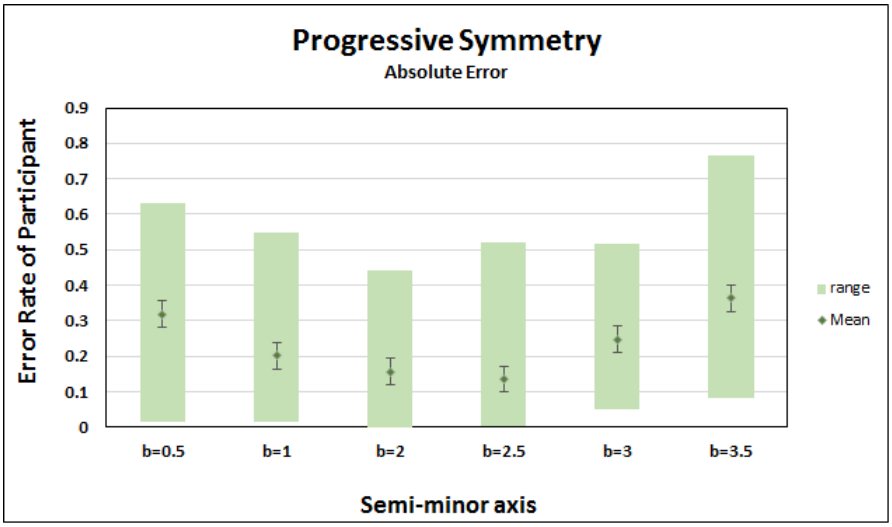


Figure 7.24: Performance analysis for task Progressive Symmetry with absolute error

Descriptive Statistics								
b_1	Mean Rank	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
0.5	4.79	0.32	0.15	0.02	0.70	0.20	0.33	0.40
1	2.84	0.20	0.12	0.02	0.48	0.08	0.20	0.30
2	2.39	0.16	0.12	0.00	0.52	0.05	0.13	0.22
2.5	1.84	0.14	0.09	0.00	0.45	0.06	0.12	0.20
3	3.80	0.25	0.13	0.05	0.58	0.15	0.23	0.35
3.5	5.34	0.36	0.15	0.08	0.70	0.27	0.35	0.45

Table 7.23: Descriptive Analysis of task Progressive Symmetry with Mean Rank, Minimum, Maximum and Percentiles

7.8.1.2 Friedman Analysis

We further support our observations deduced from the graph using Friedman test analysis, which suggests there is a statistically significant difference in error rate depending on the levels of independent variable i.e. length of the semi-minor axis of the ellipse, in correlation estimation, ($\chi^2(5) = 97.013, p = 0.000$). The descriptive analysis of this task in given in Table 7.23. The Friedman mean-ranking (see Table 7.23) further gives us an estimate of the ordering of independent levels based on the accuracy of user estimation with $b = 2.5$ being the most optimum level as it has the lowest rank and thus the least average error value.

7.8.1.3 Wilcoxon Test

Further, we perform the post-hoc test of *Wilcoxon signed-rank* analysis, to check which groups are in particular different from each other. We perform pairwise comparison of

all independent levels with the reference independent level ($b = 2.5$). The Wilcoxon signed-rank test shows the following results:

- A change in independence level from $b = 2.5$ to $b = 2$ did not elicit a statistically significant change in accuracy of correlation estimation ($Z = -1.958, p = 0.050$). Indeed, median error at both the levels were 0.1227 and 0.1333 respectively (Table 7.23, which are not very different from each other.
- A change in independence level from $b = 2.5$ to $b = 1$ elicits a statistically significant change in accuracy of correlation estimation ($Z = -3.448, p = 0.001$). In fact out of the two, user accuracy is better for the former.
- A change in independence level from $b = 2.5$ to $b = 0.5$ elicits a statistically significant change in accuracy of correlation estimation ($Z = -4.791, p = 0.000$). In fact out of the two, user accuracy is better for the former.
- A change in independence level from $b = 2.5$ to $b = 3$ elicits a statistically significant change in accuracy of correlation estimation ($Z = -4.717, p = 0.000$). In fact out of the two, user accuracy is better for the former.
- A change in independence level from $b = 2.5$ to $b = 3.5$ elicits a statistically significant change in accuracy of correlation estimation ($Z = -5.012, p = 0.000$). In fact out of the two, user accuracy is better for the former.

7.8.2 Raw Error Analysis

Figure 7.25 shows the graph plotted with raw error values of the participant data. We observe an inverted U-shaped curvilinear pattern and most part of the range for each level lies in negative y-axis which suggests that for all progressive symmetry levels, the participants generally tend to underestimate correlation values. However, this underestimation is the highest at the edges where progressive symmetry levels are the highest, as compared to the average error value in the middle where there is absence of progressive symmetry.

7.8.3 Summary

From this task, we can conclude that as the progressive symmetry moves away from the optimum value in either direction (increase or decrease in semi-minor axis), there tends to be a corresponding increase in error rate. Thus, the **human perception of correlation in scatter plots is affected by a change in progressive symmetry level** which in turn decrease the accuracy of user's estimate of correlation.

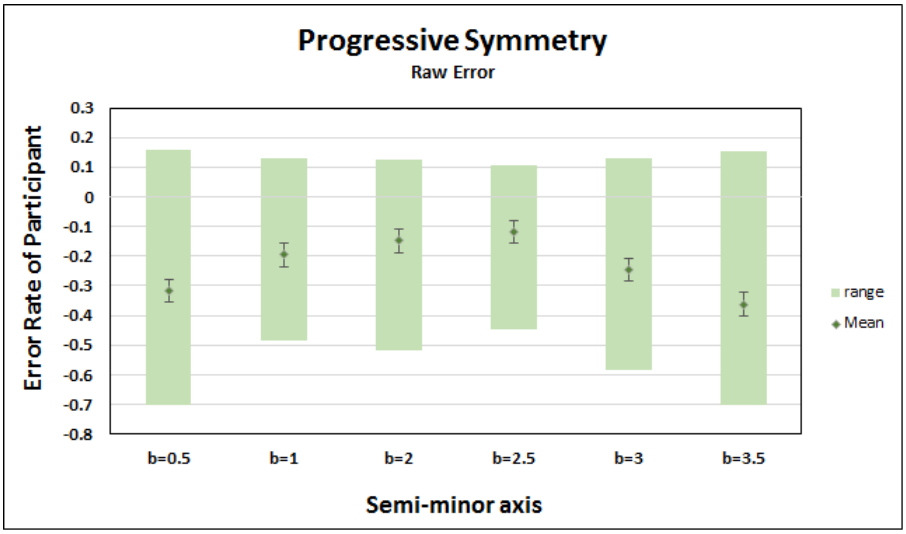


Figure 7.25: Performance analysis of task Progressive Symmetry with raw error value

7.9 Subjective Result Analysis

After the end of the main experiments, participants were asked to complete a feedback survey form which assessed their ease-of estimation-level for different scatter plots categories using three multiple choice questions as explained in Section 5.6. The feedback form is given in Appendix B for reference. The three questions evaluated the ease of estimating correlation for a) positive vs. negative (Figure 7.26), (b) uniform vs. non-uniform (Figure 7.27) and c)low density vs. high density scatter plots (Figure 7.28) respectively.

Considering all the 35 participants' responses, we can easily conclude that its easier to estimate correlation for uniform scatter plots having the standard elliptical shape compared to non-uniform ones presented during the experiment. This is justified in accordance with the statistics obtained from the tasks Distribution, Progressive Symmetry and Reflective Asymmetry which reveals a high error rate for the non-uniform patterns. Further, most of the participants (66%) find no difference in estimating positively correlated scatter plots compared to negatively correlated scatter plot. Lastly, a majority of participants find its easier to estimate correlations when a larger number of data points are used to plot the scatter plot. Only a small proportion (14%), believes that there is no major difference between using either desnsity values for plotting scatter plots which further solidifies our result from the objective analysis of task Density that reveals that average error of user's estimation of correlation tend to be highest when scatter plot density is the lowest.

Thus, the subjective analysis is in accordance with the objective analysis.

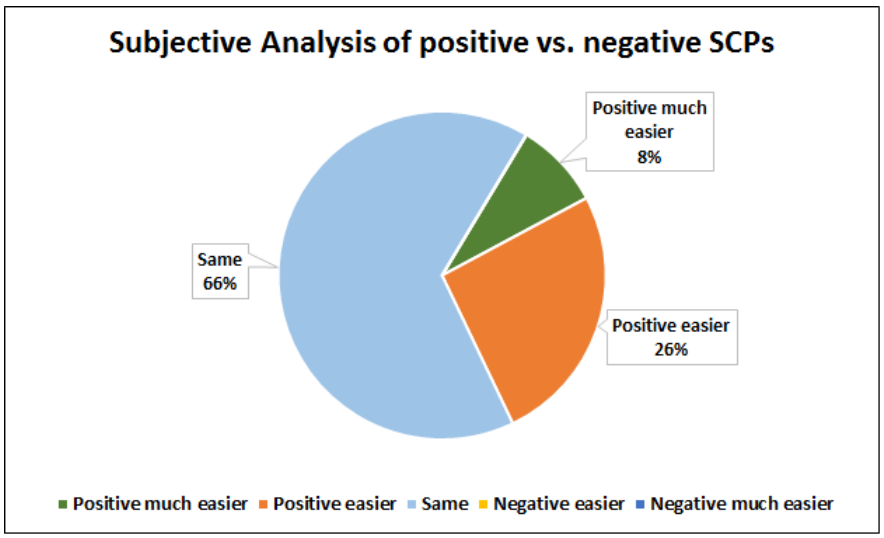


Figure 7.26: Participant's ease-of-estimation rating for Positive vs. Negative scatter plots

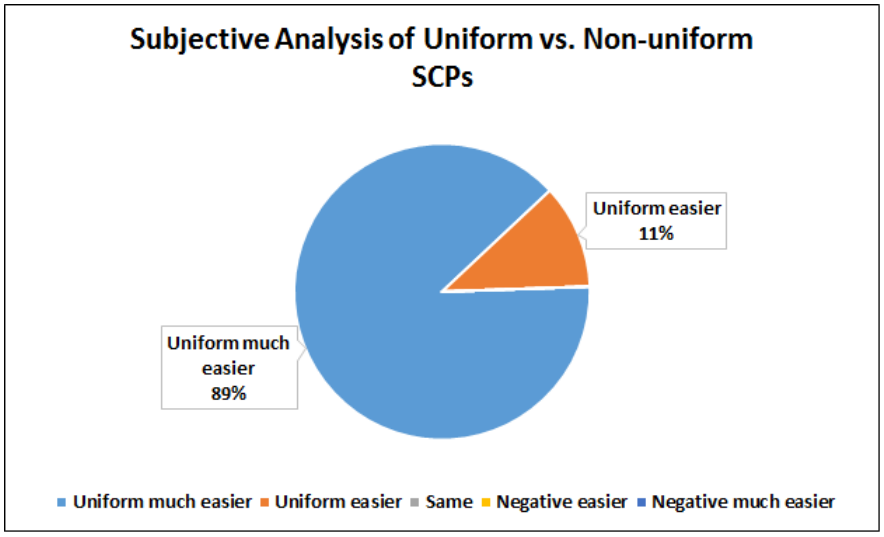


Figure 7.27: Participant's ease-of-estimation rating for Uniform vs. Non-Uniform scatter plots

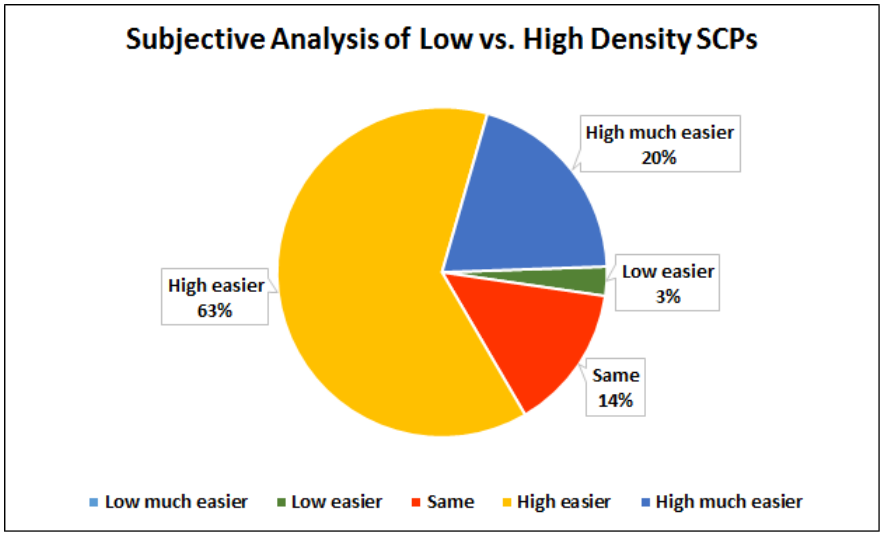


Figure 7.28: Participant's ease-of-estimation rating for Low vs. High Density scatter plots

Chapter 8

Conclusion

The goal of our research was to study the user perception of correlation in context of scatter plots. We studied various design parameters that have an impact on user perception while estimating correlations in scatter plots.

The successful completion of the study revealed the findings summarized in Section 8.1. Section 8.2 contains some discussions regarding the application of these finding. Towards the end, we mention directions for possible future work in Section .3.

8.1 Summary

In this project, we conducted an empirical study to observe variation in user performance in correlation estimation task based on different physical design features of the scatter plots. We observe the effect of variation in number of data points used to plot the scatter plot (density) and point-cloud shape of scatter plot with respect to distribution, reflective asymmetry and progressive symmetry. We measure the impact of each of these factor in terms of average of participant error. We also study the applicability of the Weber's law to measure "accuracy" and "precision" of correlation estimation in scatter plots.

On a whole, the results reveal that the human judgment of correlation is negatively impacted by variation in distribution, progressive symmetry and reflective asymmetry and a lower accuracy level is obtained for low density scatter plots. We also falsify the claims of previous studies done by Harrison *et al.* (2014) which states that Weber's law can be applicable to model the "precision" and "accuracy" of user estimates of correlation in scatter plots. The results also suggest that the value of JND in correlation is dependent on the reference correlation chosen. Further, subjective analysis of participants' response confirms that a non-uniform multi-clustered point cloud shape, such as those in case of task Distribution, Progressive Symmetry and Reflective Asymmetry, is comparatively more difficult to predict as compared to the

uniform elliptical scatter plot. Also, the users preferred a scatter plot with sufficiently high density to be able to effectively understand the cloud shape.

8.2 Discussions

This research clearly proves that the human perception of correlation is affected by varying design parameter regardless of the statistical expertize of the person. There is substantial evidence of misleading visual features of scatter plots such as the point cloud shape and density due to which the estimates of correlation may be wrongly judged. The visual properties of the display may often overpower the reasoning power build over years of formal statistical training. We claim to disregard the common notion that '*given a scatter plot, correlation can be easily perceived from it by anyone with just a little training.*'

The results of our study explains the need of a thorough understanding of the design of scatter plot in addition to the "purpose" of the correlation. If the sheer purpose of the correlation is for statistical analysis, we can do away with the need of presenting a scatter plot and report just the numerical estimator. However, if the purpose of the scatter plot is to comprehend relational patterns between variables, a numerical value of correlation coefficient won't suffice as it would be very sensitive to outliers. It would require a well-designed and informative scatter plot, designed with specifications evolved during previous studies (display of regression line (Meyer and Shinar, 1992)) and ours (use of sufficiently large number of data points) so that the intuitive correlation estimates aren't significantly different from the actual correlation coefficient.

A majority of the statistical software graphics packages available in the market today, fail to generate any additional information with the scatter plot that is relevant and focuses on the association between the variables tested. As a result, the viewer may be devoid of the exact amount of impact the outliers may have on the correlation value. Similarly, a numerical quantity is insufficient to detect partial correlations present in the data set. Thus, we believe that user perception is unreliable and so is the statistical coefficient of correlation. Neither one of the two can exist without the other. It may be in the best interest of viewers, if both the scatter plots and correlation coefficient were displayed in combination with each other so that human cognitive functions could be interspersed with the statistical training received to render highest level of scatter plot understanding. Doherty *et al.* (2007) claims the two of them to be "the complementary sources of intuitions about relationships" which shouldn't be interpreted alone.

8.3 Future Work

Lew rightly said, there is no area in which we have enough data (Lewandowsky and Spence, 1989), although we believe our research gives fair amount of knowledge about the different metrics people employ for comprehending scatter plots. The study verifies that cloud shape is an examinatory factor for correlation perception, however, there remains much more to be learned.

Future works could be undertaken to study other design parameters such as size, shape and color of the data point plotted on the scatter plot, delving into the cognitive aspect of perception. The effect of the variations of slope of the point cloud and the slope of the regression line could also be taken into consideration to determine any variations in user perception at any particular slope.

The scope of our study was limited to scatter plots only as they are perceived to be one of the most powerful visualization which has sustained over the years. However, it would be worth evaluating in future the perception of correlation using other visualizations such as bar charts, parallel co-ordinates, or line graphs. In addition, it could be applied to various visualization tasks such as outlier detection, clustering and classification tree generation. Also, the research could be extended by studying multidimensional data instead of bivariate data so as to gather more generalized result for any data type.

Due to time constraint, our study covered a limited number of point cloud designs, focusing on clouds along the diagonals i.e. $45°$ and $-45°$ lines. We believe future works could be undertaken to investigate user perception for point clouds lying on horizontal or vertical lines passing through the centre of the scatter plot. Also, task JND reports that the JND value for correlation is dependent on reference correlation chosen. Hence, it would be interesting to see whether much smaller differences between two scatter plots ($d < 0.05$) can be effectively deduced from some particular reference correlation point.

On a concludary note, we need to direct future works towards deriving an exhaustive class of properties that not only can define variations in user perception but does so with high degree of precision. There is a lot that has been theoretically proven previously but lacks the underlying foundation and level of confidence an empirical study provides. We believe, this research would act as a stepping stone for generating graphs that maximize comprehensibility, aids decision making and prove beneficial for the Information Visualization community.

Appendix A

Pre-study Presentation

Scatter Plots

Empirical Study to observe visual perception of Correlation

Aim of the study

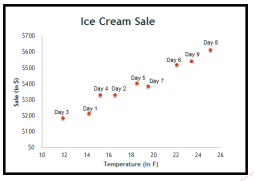
- To study how humans may use their visual perception to judge *Correlation* (henceforth denoted by symbol R) just by looking at the *Scatter Plots.*
- Correlation is a statistical measure that indicates how strongly pairs of variables are related. It can take a value between -1 and +1 where R=1 depicts perfect Positive Correlation and R=-1 depicts perfect Negative Correlation.
- The aim of this experiment is to observe how people judge the correlation value when shown different Scatter Plots.
- The answers are not expected to be exact but we would like you to estimate Correlation as close as possible to what will be generated by Mathematical formula.

What are Scatter Plots?

- A scatter plot is a graph in which two variables (X and Y) are plotted as ordered pairs in a coordinate plane.
- The more closer the points, the more amount of correlation present in the data set.

Example 1: Let us consider 2 variables i.e. X and Y where X denotes the Temperature (in F) and Y denotes the total amount earned from sale of Ice-creams (in $).

Observation Day	Temperature (X)	Sale (Y)
Day 1	14.2	$215
Day 2	16.4	$325
Day 3	11.9	$185
Day 4	15.2	$332
Day 5	18.5	$406
Day 6	22.1	$522
Day 7	19.4	$412
Day 8	25.1	$614
Day 9	23.4	$544

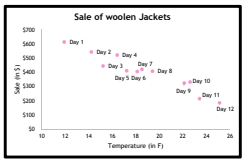


Example 2: Let us consider 2 variables i.e. X and Y where X denotes the Height (in cm) and Y denotes the size of the Shoe.

Customer	Height (in cm)	Shoe size
Customer 1	185	6.5
Customer 2	153	4.5
Customer 3	157	5
Customer 4	160	5.5
Customer 5	165	5
Customer 6	172	5.5
Customer 7	180	6
Customer 8	175	6
Customer 9	162	5
Customer 10	190	7

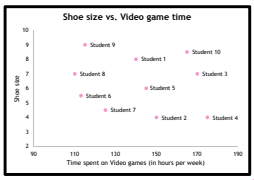
Example 3: Let us consider 2 variables i.e. X and Y where X denotes the Temperature (in F) and Y denotes the sale of Woolen jackets.

Observation Day	Temperature (X)	Sale (Y)
Day 1	11.9	$614
Day 2	14.2	$544
Day 3	15.2	$445
Day 4	16.4	$522
Day 5	17.2	$412
Day 6	18.1	$406
Day 7	18.5	$421
Day 8	19.4	$408
Day 9	22.1	$325
Day 10	22.6	$332
Day 11	23.4	$215
Day 12	25.1	$185



Example 4: Let us consider 2 variables i.e. X and Y where X denotes the amount of time spent on Video games (in hours per week) and Y denotes the shoe size.

Student ID	Time spent (X)	Shoe size (Y)
Student 1	140	8
Student 2	150	4
Student 3	170	7
Student 4	175	4
Student 5	145	6
Student 6	113	5.5
Student 7	125	4.5
Student 8	110	7
Student 9	115	9
Student 10	165	8.5



How can you deduce Correlation from Scatter Plots?

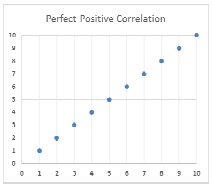
- **Positive Correlation**: as one variable increases so does the other & hence we get a Scatter Plot that has a straight line gradually rising **upwards** from left to right.
- Example:
  - The more money you save, the greater savings you will have.

  - As the child grows, his clothing size also increases.

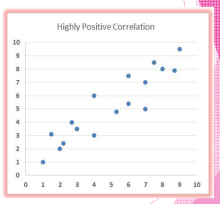
  - People are likely to earn more compound interest if they invest their money for longer time durations.

  - The more education years people have completed, the higher is their potential to earn better.

Positive Correlation



Correlation = 1.0



Correlation = 0.873

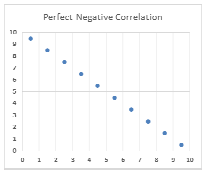
How can you deduce Correlation from Scatter Plots?

- **Negative Correlation**: as one variable increases, the other decreases other & hence we get a Scatter Plot that has a straight line that rises **downwards** from left to right.
- Example:
  - The more a person's expenditure increases, the lesser savings he is likely to have.

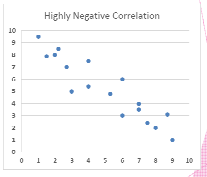
  - As amount of rain increases, the friction in car tyres decreases.

  - The more that people are vaccinated for a specific illness, the less that illness occurs.

Negative Correlation



Correlation = -1



Correlation = -0.87

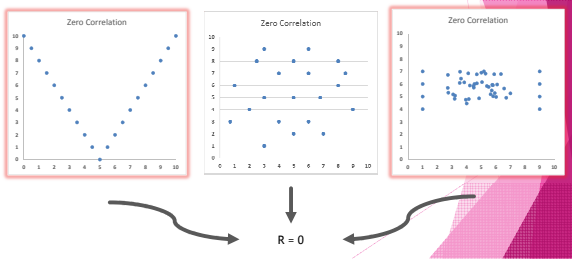
How can you deduce Correlation from Scatter Plots?

- **Zero Correlation**: there is no apparent relationship between the variables & hence we get a Scatter Plot that has data points scattered all over.
- For example: Time spent on Video games and shoe size appear to have no correlation; as one increases, the other has no effect

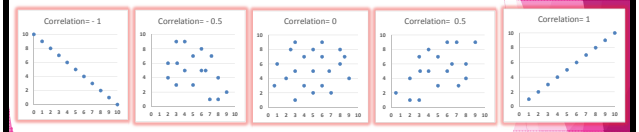
Correlation = 0

Mapping between Scatter Plots and Correlation

Many -to- one mapping: It means different Scatter Plots can correspond to one Correlation value but there is only 1 specific Correlation value for each Scatter Plot.

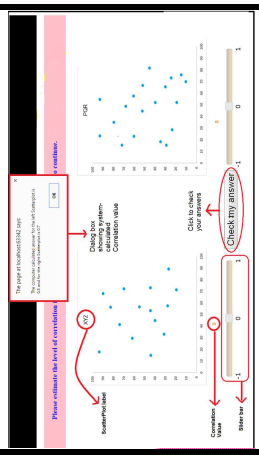


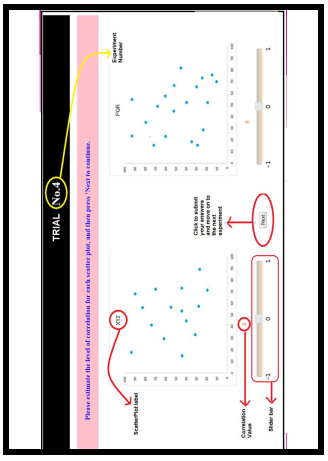
Summary



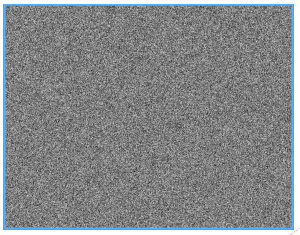
Itinerary for today.

- 5 Training trials for Scatter Plots.
- 20 Testing trials for Scatter Plots.
  - Break for at least 2 minutes.
- 25 Testing trials for Scatter Plots.
  - Break for at least 4 minutes.
- 25 Testing trials for Scatter Plots.
  - Break for at least 2 minutes.
- 25 Testing trials for Scatter Plots.
- Emailing the data file.
- Feedback Survey form.





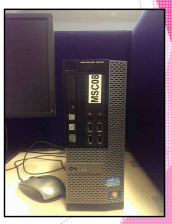
- After each experiment, there is display of a blank gray-scale screen for 2 seconds. This is called Masking effect in psychology field and is essential for relaxing the eye vision before moving onto the next experiment number.



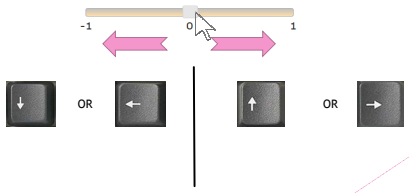
General instructions.

- At any stage, you can't go back to change your answer to the previous question. So please click the Next button only when fully satisfied with your answer.
- Please clarify all your doubts during the Training session itself. Moderators normally cannot assist once the Testing session starts.
- There is no time limit to answer a question, so you can take as long as you want to ponder upon the question.
- Please judge the value of Correlation for a Scatter Plot based solely on that particular Scatter Plot ONLY (irrespective of the Scatter Plot displayed along side it).
- Please ignore the label at top of the Scatter Plot. That is used as ID of Scatter Plot.
- The 2 Scatter Plots are randomly placed, meaning it isn't necessary that Correlation of left one is always greater than the right one or vice versa. Either of them could be greater or smaller & in some cases both could be equal.

Your user number will be the system number you are sitting on. It is written on the CPU (MSC # # #). Please ensure you enter the correct User Number.



Move the slider bar left or right with the mouse to choose the appropriate correlation value. To ease things up, you can click on the slider once, and then use the Up and Down OR Right and Left arrow keys on the keyboard to move slider left or right.



Feedback Survey Form

- Enter the User ID (pasted on the CPU in front of you) on the feedback form.
- Answer the 3 multiple-choice questions in the feedback form (select only 1 answer for each question).
- Leave the completed feedback form next to the computer used to take the study.

Thank you ☺

Questions???

Appendix B

Feedback Survey Form

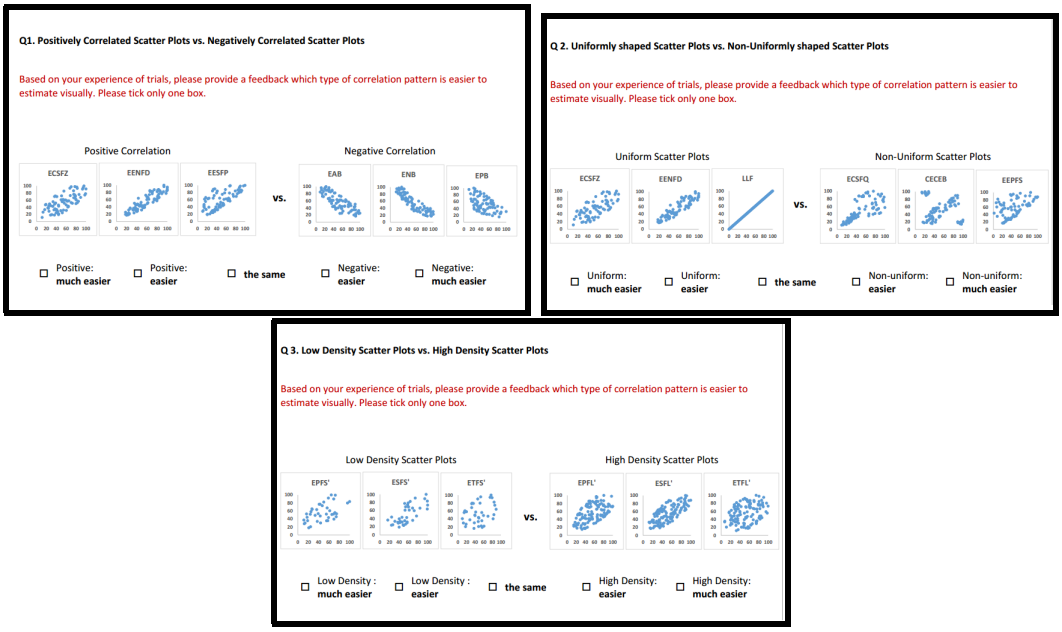


Figure B.1: Feedback survey form with three multiple choice questions

Appendix C

Measurement for generation of scatter plots

Database for measurement used in Scatter Plot generation											
S.No.	Secret Code	Type	Data points	Distribution	Expected R	Program generated R	Ellipse1 semi-major axis (a)	Ellipse1 semi-minor axis (b_1)	Circle radius (r)	Ellipse2 semi-minor axis (b_2)	Session Type
1	LLF	Straight line	80	NULL	1	1	NULL	NULL	NULL	NULL	testing
2	LLB	Straight line	80	NULL	-1	-1	NULL	NULL	NULL	NULL	testing
3	ENDF	Ellipse	80	NULL	0.95	0.950218	7	1.5	NULL	NULL	testing
4	ENF	Ellipse	80	NULL	0.9	0.900652796	7	2	NULL	NULL	testing
5	EADF	Ellipse	80	NULL	0.85	0.850277	7	2.15	NULL	NULL	testing
6	EAF	Ellipse	80	NULL	0.8	0.802212	7	2.25	NULL	NULL	testing
7	ESF	Ellipse	80	NULL	0.7	0.705470091	7	2.5	NULL	NULL	testing
8	ECDF	Ellipse	80	NULL	0.65	0.65344599	7	2.6	NULL	NULL	testing
9	ECF	Ellipse	80	NULL	0.6	0.60890257	7	2.7	NULL	NULL	testing
10	EPDF	Ellipse	80	NULL	0.55	0.559183953	7	2.75	NULL	NULL	testing
11	EPF	Ellipse	80	NULL	0.5	0.502779217	7	3	NULL	NULL	testing
12	EQDF	Ellipse	80	NULL	0.45	0.451242548	7	3.25	NULL	NULL	testing
13	EQF	Ellipse	80	NULL	0.4	0.409464445	7	3.5	NULL	NULL	testing
14	ETDF	Ellipse	80	NULL	0.35	0.350612732	7	3.75	NULL	NULL	testing
15	ETF	Ellipse	80	NULL	0.3	0.308742326	7	4	NULL	NULL	testing
16	EDF	Ellipse	80	NULL	0.2	0.203332	7	5	NULL	NULL	testing
17	EEDF	Ellipse	80	NULL	0.15	0.150178	7	5.25	NULL	NULL	testing
18	EEF	Ellipse	80	NULL	0.1	0.10036011	7	5.5	NULL	NULL	testing
19	EZDF	Ellipse	80	NULL	0.05	0.0504634	7	6	NULL	NULL	testing
20	EZF	Ellipse	80	NULL	0	0.00166495	7	7	NULL	NULL	testing
21	EEB	Ellipse	80	NULL	-0.1	-0.102951327	7	5.5	NULL	NULL	testing
22	ETB	Ellipse	80	NULL	-0.3	-0.308317358	7	4	NULL	NULL	testing
23	ETDB	Ellipse	80	NULL	-0.35	-0.351035464	7	3.75	NULL	NULL	testing
24	EQB	Ellipse	80	NULL	-0.4	-0.4025075	7	3.5	NULL	NULL	testing
25	EQDB	Ellipse	80	NULL	-0.45	-0.450895691	7	3.25	NULL	NULL	testing
26	EPB	Ellipse	80	NULL	-0.5	-0.500532066	7	3	NULL	NULL	testing
27	EPDB	Ellipse	80	NULL	-0.55	-0.547217043	7	2.75	NULL	NULL	testing
28	ECB	Ellipse	80	NULL	-0.6	-0.60045	7	2.7	NULL	NULL	testing
29	ESB	Ellipse	80	NULL	-0.7	-0.706981714	7	2.5	NULL	NULL	testing
30	EAB	Ellipse	80	NULL	-0.8	-0.80945	7	2.25	NULL	NULL	testing
31	ENB	Ellipse	80	NULL	-0.9	-0.896141524	7	2	NULL	NULL	testing
32	ESFL	Ellipse	100	NULL	0.7	0.715095205	7	2.5	NULL	NULL	testing
33	ESFLL	Ellipse	120	NULL	0.7	0.709400004	7	2.5	NULL	NULL	testing
34	ESFS	Ellipse	60	NULL	0.7	0.700870643	7	2.5	NULL	NULL	testing
35	ESFSS	Ellipse	40	NULL	0.7	0.701744526	7	2.5	NULL	NULL	testing
36	EPFSS	Ellipse	40	NULL	0.5	0.504870117	7	3	NULL	NULL	testing
37	EPFSS	Ellipse	60	NULL	0.5	0.508648617	7	3	NULL	NULL	testing
38	EPFL	Ellipse	100	NULL	0.5	0.503858749	7	3	NULL	NULL	testing
39	EPFLL	Ellipse	120	NULL	0.5	0.517633103	7	3	NULL	NULL	testing
40	ETFLL	Ellipse	120	NULL	0.3	0.306645409	7	4	NULL	NULL	testing
41	ETFL	Ellipse	100	NULL	0.3	0.304129359	7	4	NULL	NULL	testing
42	ETFS	Ellipse	60	NULL	0.3	0.307671463	7	4	NULL	NULL	testing
43	ETFSS	Ellipse	40	NULL	0.3	0.303690845	7	4	NULL	NULL	testing
44	CECNB	Circle + Ellipse+Circle	80	30:20:30	-0.9	-0.902518208	7	3	1	NULL	testing
45	CECSB	Circle + Ellipse+Circle	80	25:30:25	-0.7	-0.701300913	7	3	1	NULL	testing
46	CECPB	Circle + Ellipse+Circle	80	20:40:20	-0.5	-0.505626416	7	3	1	NULL	testing
47	CECTB	Circle + Ellipse+Circle	80	15:50:15	-0.3	-0.300781894	7	3	1	NULL	testing
48	CECEB	Circle + Ellipse+Circle	80	10:60:10	-0.1	-0.104187874	7	3	1	NULL	testing
49	CECDDF	Circle + Ellipse+Circle	80	5:70:5	0.25	0.251415715	7	3	1	NULL	testing
50	CECPF	Circle + Ellipse+Circle	80	0:80:0	0.5	0.505937987	7	3	1	NULL	testing
51	EENFDD	Big + Small Ellipse	80	NULL	0.9	0.901554211	7	2.5	NULL	1.5	testing
53	EENFD	Big + Small Ellipse	80	NULL	0.9	0.908548668	7	2	NULL	2	testing
54	EENFT	Big + Small Ellipse	80	NULL	0.9	0.900370077	7	3	NULL	1	testing
55	EESFQ	Big + Small Ellipse	80	NULL	0.7	0.702204028	7	4	NULL	1.5	testing
56	EESFDD	Big + Small Ellipse	80	NULL	0.7	0.701600472	7	2.5	NULL	2.5	testing
57	EESFP	Big + Small Ellipse	80	NULL	0.7	0.70196274	7	5	NULL	1	testing
58	EEPFP	Big + Small Ellipse	80	NULL	0.5	0.501744488	7	5	NULL	2	testing
59	EEPFT	Big + Small Ellipse	80	NULL	0.5	0.502051248	7	3	NULL	3	testing
60	EEPFS	Big + Small Ellipse	80	NULL	0.5	0.501569054	7	7	NULL	1	testing
61	ECSFQ	Ellipse+Circle	80	0:40:40	0.7	0.706781962	7	1	4	NULL	testing
62	ECSFT	Ellipse+Circle	80	0:40:40	0.7	0.705585807	7	2	3	NULL	testing
63	ECSFZ	Ellipse+Circle	80	0:40:40	0.7	0.70590253	7	2.5	0	NULL	testing
64	ECSFE	Ellipse+Circle	80	0:40:40	0.7	0.709058166	7	3.5	1	NULL	testing
65	ECSFD	Ellipse+Circle	80	0:40:40	0.7	0.704765374	7	3	2	NULL	testing
66	ECSFP	Ellipse+Circle	80	0:40:40	0.7	0.704649803	7	0.5	5	NULL	testing
67	LLF	line	80	NULL	1	1	NULL	NULL	NULL	NULL	training
68	LLB	line	80	NULL	-1	-1	NULL	NULL	NULL	NULL	training
69	EEZZEE	zero correlation	80	NULL	0	0.00051	7	7	NULL	NULL	training
70	EPB	ellipse	80	NULL	-0.5	-0.49	7	3	NULL	NULL	training
71	EPF	ellipse	80	NULL	0.5	0.504	7	3	NULL	NULL	training
72	ESF	ellipse	80	NULL	0.7	0.707890293	7	2.5	NULL	NULL	training
73	ETB	ellipse	80	NULL	-0.3	-0.297327292	7	4	NULL	NULL	training
74	OOEF	odd shape (2 CIRCLES)	80	NULL	0.1	0.109245261	NULL	NULL	1	NULL	training
75	OOPB	odd shape (LINE + CIRCLE)	80	NULL	-0.5	-0.49139235	7	0.5	1	NULL	training
76	ETBETB	ellipse	80	NULL	-0.3	-0.308116923	7	4	NULL	NULL	training

Table C.1: Database for measurement used in Scatter Plot generation

Appendix D

Software Implementation

D.1 Software Implementation

We implement the software program mainly in HTML, using JavaScript for client side scripting, PHP for server side scripting and CSS for defining the design and layout of the web page. We explain each one in detail in following sub-sections.

D.1.1 JavaScript

JavaScript is a client-side scripting language. In simple terms, it is useful for making interactive web pages, validating user input, interacting with local storage (cookies & browser storage), etc.

In our software we use 5 main JavaScript files *json2.js, jstorage.js, jquery.js, jquery-ui.js,* and *jstorage.min.js.* The usage and working of each is explained below.

D.1.1.1 json2.js

JSON is short for JavaScript Object Notation. It is a convenient way to store information in an organized, easy-to-access manner. It renders human-readable collection of data that we can access in a really logical manner. The *json2.js* creates a global *JSON* object containing *stringify* method which converts a JavaScript value to a JSON string. Its syntax is: *JSON.stringify(value, replacer, space).*

value : any JavaScript value, usually an object or array, to be converted to a json string.

replacer (optional) : an optional parameter that determines how object values are stringified for objects. It can be a function or an array of strings. It serves as a whitelist for selecting the properties of the value object to be included in the JSON string.

space (optional) : a String or Number object that's used to insert white space into the output JSON string for readability purposes.

We *stringify*, all the participant's data into a *json* variable which is then used as a parameter to create a *blob* object using Blob() constructor. A Blob object consists of the concatenation of the array of values given to it as parameter. We create a URL to the json data using that blob which is then available for downloading by the user. Each downloadable json file acts as a backup copy of the participant's data and is uniquely identifiable since it is named using userID of the participant such as 'MSC777backup.json' and 'MSC20backup.json'. The files are downloaded on individual computer systems from where they are emailed to a backup email account for storage purpose.

D.1.1.2 jstorage.js

It implements a simple local storage wrapper to save '*key-value*' database on the browser side. It supports storing Strings, Numbers, JavaScript objects, Arrays and even native XML nodes.

We make use of 3 main functions defined in the script: *jStorage.set()*, *jStorage.get()* and *jStorage.index()* which are described below:

jStorage.set(): It's syntax is - *set(key, value[, options])*. It is used to save a value to local storage.

- *key*- needs to be string otherwise an exception is thrown;
- *value*- can be any JSONeable value, including objects and arrays or a XML node
- *options* (optional)- the only available option is *options.TTL* which can be used to set the TTL (Time-to-live) value to a *key*.

jStorage.get(): It's syntax is - *get(key[, default])*. It is used to retrieve the value of the *key* if it exists else returns the *default* value.

jStorage.index(): It returns all the *keys* currently in use as an array.

We use the *set* function to store 100 records for each participants during the training (5 records) and main trials (95 records). Each record is made of 29 key-value pairs of information: *USER NO, AGE, GENDER, OCCUPTION, COLOR-BLIND, FAMILIARITY, SESSION TYPE, QUESTION NO, STIMULI NO, LEFT SCATTERPLOT, ESTIMATED VALUE1, ACTUAL VALUE1, MATCH1, DIFF1, OVERESTIMATE1, UNDERESTIMATE1, RIGHT SCATTERPLOT, ESTIMATED VALUE2, ACTUAL VALUE2, MATCH2, DIFF2, OVERESTIMATE2, UNDERESTIMATE2, Q_GREATER, A_GREATER, MATCH, START TIME, STOP TIME* and *RESPONSE TIME*. Once we store these 29 information for a record, we append an

additional key-value pair : (*space*,";") at the end. This is used to delimit each record with a semicolon so that it is easier to segregate and transfer them to different rows in MS Excel for analysis. Thus, in totality we store an array of 30 information for each record.

Lastly, we use the *get*() to retrieve all the key values into a variable called *data* by running the *for* loop as many times as the length of the storage array containing all the record variables i.e. 'jStorage.index().length' times.

D.1.1.3 jquery.js

It is a JavaScript library which contains one line methods for commonly used tasks requiring many lines of JavaScript code to be accomplished. It have been designed specifically to simplify HTML document traversing, animation, event handling, and Ajax interactions. We prefer jQuery over JavaScript as it saves time and is easier to implement. We mainly use jQuery to:

- return values of the two sliders, representing correlations estimated by the participant.
- set values of form fields, such as START TIME and END TIME.

D.1.1.4 jquery-ui.js

It contains the code for jQueryUI library. While jQuery is the core library, jQueryUI is built on top of it. It provides abstractions for low-level interaction and animation, advanced effects and high-level, themeable widgets (slider bars, accordion, button, menu, dialog, etc) which are useful for building highly interactive web pages. It also helps style these powerful user-interface (UI) elements in this library. We mainly use jQueryUI in our software to implement/deploy attractive slider bars as means to collect user input.

The slider bars are then styled using CSS which provide functionality to adjust color, width and height of the slider bar and its handle.

Lastly using jQueryUI, we append slider labels under the slider bars to specify their range. The minimum and maximum range are specified as -1 and +1 respectively with tick marks at -1, 0 and +1 positions and incremental steps of 0.05.

D.1.1.5 jstorage.min.js

This script file has the same functionality as *jstorage.js*. It is the compressed version (whitespaces & comments stripped out, shorter variable names, etc) in order to preserve bandwidth. As it has a smaller file size, it loads faster too. While it is

recommended to use this compressed version in production environment, we nevertheless include the maxified version (*jstorage.js*) as part of our software for debugging purpose and for future references by researchers who would like to study the code.

D.1.2 HTML

We have a total of 106 HTML files, 3 for displaying clocks during breaks (*clock1.html, clock2.html, clock3.html*), 1 for presenting pre-study questionnaire to participants (*intro.html*), 1 for signaling the termination of training session and beginning of main trials (*welcome.html*), 1 for signaling the termination of experiment (*lastpage.html*), 5 for presenting Training trials (*Training_ 1.html, Training_ 2.html, Training_ 3.html, Training_ 4.html, Training_ 5.html*) and remaining 95 for presenting the main testing trials(*Testing_ 1.html, Testing_ 2.html*, etc). In the following subsections, we briefly explain the functionality of each.

D.1.2.1 intro.html

It creates a short descriptive form to capture participant detail prior to starting of the experiment. The participants must first enter their UserID in the text-box provided. This field is mandatory and cannot be left blank. Unless the participant has filled it, the form wont be submitted and participant cannot proceed with the study. Next the participants must provide details of their age group, gender, occupation and color-blindness. To gather user response for each of them, we provide radio buttons instead of text boxes as they are much easier to use. Lastly, their familiarity with scatter plots is recorded using a five level Likert Scale.

D.1.2.2 clock1.html, clock2.html, clock3.html

The software implements three identical clocks which differ only by the amount of time the break is scheduled for. *clock1* and *clock3* offer a 2 minute break to participants whereas *clock2* offers a longer break of 4 minutes as explained in Section 5.4.5. During the break, the participants are presented with a pie clock on the screen which shows the break time remaining. There is a 'Continue' button which has been disabled to ensure participants engage in a break. Once the pie clock has vanished completely, the participants may continue the study by clicking the 'Continue' button which is then enabled or may extend their break time as per their convenience.

D.1.2.3 welcome.html

This web page acts as a demarcation between the 5 training session trials and the remaining 95 main trials. It specifies to the participants the end of training session

and provides them with a 'Continue' button, which when clicked begins the main trials.

D.1.2.4 training_1.html, training_2.html, etc

Each of these files have dual functionality of displaying the trials to the participants and collecting their response as form elements. Each displays the training trial number top of the screen, along with 2 SCPs and the corresponding slider bars. The updated values from slider bar are reflected on the screen in a text box, between the scatter plot and its slider bar. Each of these HTML files have values pre-fixed for the following form elements- *corr1, corr2, question, stimuli, session_type, left_SCP and right_SCP* based on the 2 stimuli displayed in that particular trial. In all these HTML files, *session_type* is set to the value 'Training' and *question* & *stimuli* have identical values with each starting from 1 and going up to 5 in a numerically increasing order. The values of rest of the form elements are either entered by the user or calculated by the software using the values entered by the user.

The values of *slider1* and *slider2* are entered by the user. The software calculates *Q_Greater* and *A_Greater* using these in addition to values of *corr1* & *corr2*. If *corr1* is greater than *corr2*, the value of *Q_Greater* is 'Left' and if its smaller, the value is 'Right'. In case of the two being equal, value is set as 'Equal'. In a similar manner, value of *A_Greater* is calculated using *slider1* and *slider2*. Using *Q_Greater* and *A_Greater*, the value of *Match_Greater* is derived. It's value is 'Yes' if the two are equal else 'No'. Similarly, *response* time is calculated using difference of *start* time and *end* time.

Additionally, if *slider1* equals *corr1*, the value of *match1* is 'Yes' else 'No'. Subtracting *corr1* from *slider1* gives the value of *diff1*. The sign (positive or negative) of *diff1* signifies the value of *underestimate1* and *overestimate1*. If it is positive, value of *overestimate1* is 'Yes' and *underestimate1* is 'No'. The values are reversed if the sign is negative. In case, *diff1* is equal to 0, both *underestimate1* and *overestimate1* are set to value 'No'. The same procedure holds for calculating the values of its counterpart *diff2, match2, underestimate2, overestimate2*.

D.1.2.5 testing_1.html, testing_2.html, etc

The basic functionality of these HTML files remain similar to that of Training session files. The only major difference is the value of *session_type* which is modified to 'Testing'. Also, the value of *question* is no longer same as that of *stimuli* because the trials are randomized. The question numbers start from 1 and go up to 95 in numerically increasing manner whereas the stimuli have same range but the ordering is randomized as explained in Section 5.4.3.

D.1.2.6 lastpage.html

This is the concluding screen presented to the participants informing them that the experiment has ended. It has a 'Click on me to finish experiment' button. When clicked, it downloads a json file that contains all the participant's demographic data along with his responses for the trials in form of 100 records, each of which is semi-colon separated as explained in Section 6.2.1.2.

D.1.3 PHP

PHP (Hypertext Preprocessor) is a server-side scripting language for web development. Its main use is to create dynamic web page content. We implement two php files, *Tmain.php* for Training session trials and *main.php* for the Testing trials. The purpose of these php files is mainly to implement functionality for collecting user response which can be later utilized for statistical analysis.

When the user clicks on the 'Submit' button after completing the pre-study questionnaire, *Tmain.php* script runs in the background. It reads up the information submitted by the participant, such as UserID, age, occupation, familiarity, etc . Since these information remain static for a particular participant throughout the experiment, the php script creates session variables for each one of them. Thus, we have 6 session variables- *userno, age, gender, occupation, color_blind, familiarity* - which are stored once and used across multiple pages. By default, session variables last until the user closes the browser.

Henceforth, for each training session trials, the script just stores the values of the remaining 23 global variables that change with each training session trial. These include -*slider1, slider2, corr1, corr2, question, stimuli, session_type, start, end, response, match1, match 2, diff1, diff2, overestimate1, overestimate2, underestimate1, underestimate2, Q_Greater, A_Greater, Match_Greater, left_SCP* and *right_SCP*. Next, it creates a record of 29 variables per trial for storage in a text file using the 6 session variables and 23 global variables. The name of the text file is personalized with respect to the UserID and is always opened in the 'append' mode when it's written to or updated. Lastly, the page is redirected to the next trial after displaying the masking screen for 2 seconds.

main.php has the same functionality as *Tmain.php* with an added feature to handle break intervals during which instead of relocating to the next trial, php script redirects user to either of the break screens (clock1.html, clock2.html or clock3.html) as explained in Section 5.4.5. Also, once all the trials have ended, the script redirects user to *lastpage.html* which signifies end of the experiment and provides link to download file containing user response.

D.1.4 CSS

We include a CSS (Cascading Style Sheet) file, *css1.css* into out software which describes how different elements (scatter plots, slider bars, trial no, etc) are to be displayed on the web page. We use it to alter the fonts, text, colors, backgrounds, margins, and layout to render a fancier user-friendly site.

Rather than writing separate style sheets for all HTML pages, we generate one common css file, *css1.css* and use the same style sheet with multiple HTML files. It helps maintain consistency in format and layout across the entire experiment. With a single style sheet to administer, it is easier to modify and update it, as and when required, and the changes are reflected across entire software. We format the following elements using *css1.css*:

- scatter plots: to display them adjacent to each other in the centre of the screen.
- slider bars: to position them vertically below the SCPs in a common row, set the width and height of the slider bar handles and give them identical attractive colors.
- 'Next' button: to position the 'Next' button exactly in the middle of the two scatter plots for user convenience.
- 'Trial No' display: to display at the top of screen, the current trial number for user reference.
- clock and 'Continue': to display the clock and a *disabled* 'Continue' button during the break time, in the middle of the screen.
- background colors, font styles, font alignment, font weight: to define hierarchy of the web page content, enhance clarity and focus user attention on specific details.

Appendix E

Stimulus Generation

E.1 Stimuli Generation

Stimulus generation is composed of 2 main processes, data generation and scatter plot generation.

E.1.1 Data Generation

The data sets for all stimuli belonging to a particular task category are generated through programs written uniquely for each task category using C++ that generates points so that a predetermined correlation (accurate at ±0.01) and a pre-determined shape of the data point cloud is obtained. Hence, there are six main programs for each of the six task categories: *Task1_JND.cpp*, *Task2_ReflectiveAsymmetry.cpp*, *Task3_ProgressiveSymmetry.cpp*, *Task4_Distribution.cpp*, *Task5_Density.cpp* and *Task6_Weber.cpp*. The following subsections explain each of them in brief:

E.1.1.1 Task1_JND.cpp

The main aim of task JND is to observe whether participants, when shown 2 different SCPs, can detect the presence of a difference in correlation between those two. We perform this task using elliptical cloud shape since it is has lot of scope for variation in shape with corresponding change in correlation value.

We use this same program to generate data set for the sub tasks of task JND i.e. task JND Coarse and task JND Fine. For task JND Coarse, we generate data set for R= [0.1, 0.5, 0.7, 0.3, 0.9, 0.8, 0.2, 0.15, 0.95, 0.05 and 0.85]. These SCPs are then paired as explained in Chapter 4. Similarly, for task JND Fine 0.7, JND Fine -0.7, JND Fine -0.3 and JND Fine 0.3, we generate data set for R= [0.7, 0.65, 0.6, 0.55, 0.45, 0.4 and 0.35], [-0.7, -0.65, -0.6, -0.55, -0.45, -0.4 and -0.35], [-0.3, -0.35, -0.4, -0.45, -0.55, -0.6 and -0.65] and [0.3, 0.35, 0.4, 0.45, 0.55, 0.6 and 0.65] respectively.

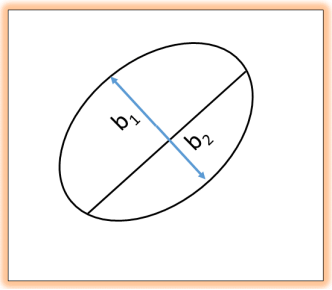


Figure E.1: Point cloud shape for task Reflective Asymmetry

The following steps generate data points for SCP having correlation R. Table in Appendix C gives value of a and b corresponding to a positive R ($R > 0$).

1. Generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b$ and $-b$.

2. Check to see if (x,y) satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= b and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 < 0$

3. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45° and add it to the set of data points else discard.

4. Repeat step 1 till the list has 80 data points in the set.

5. Calculate correlation for the 80 data points. If it matches the desired correlation R, output the data set else go to step 1.

In case of generating data points for a SCP with negative R (R<0), change the angle of rotation to -45°, while rest of the program remains the same.

E.1.1.2 Task2_ReflectiveAsymmetry.cpp

The main aim of task Reflective Asymmetry is to observe how perception of participant's vary with a change in reflective asymmetry of the cloud shape in scatter plots. All the trials of this task have a cloud shape composed of 2 half ellipses with semi-minor axes (b_1 and b_2) sharing a common major axis (a). Variation in reflective asymmetry is achieved by altering the two semi-minor axes. Figure E.1 shows one such scatter plot.

The following steps generate data points for SCP having correlation R. Each R displays 3 variations in symmetry obtained by altering values of b_1 and b_2. Table in Appendix C gives different value of a, b_1 and b_2 corresponding to a single R.

1. Generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b_1$ and $-b_1$.

2. Check to see if y >= 0 AND (x,y) satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= b_1 and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b_1^2} - 1 < 0$

3. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45° and add it to the set of data points else discard.

4. Repeat step 1 till the list has 40 data points in the set. Thus we have generated upper half of the cloud shape.

5. Again generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b_2$ and $-b_2$.

6. Check to see if y < 0 AND (x,y) satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= $b1$ and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b_2^2} - 1 < 0$

7. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45° and add it to the set of data points else discard.

8. Repeat step 5 till the list has another 40 data points in the set.

9. Calculate correlation for the 80 data points. If it matches the desired correlation R, output the data set else go to step 1.

The same program is used to generate data set for all the trials of Task Reflective Asymmetry. Hence, we use the program to generate 3 SCPs each for R= 0.3, 0.5 and 0.7.

E.1.1.3 Task3_ProgressiveSymmetry.cpp

The main aim of task Progressive Symmetry is to observe how perception of participant's vary when the scatter plot has progressive symmetric cloud shape. All the trials of this task has a cloud shape of scatter plot that is composed of an ellipse at an angle of 45° and a circle overlapping the top of the ellipse. Figure E.2 shows one such scatter plot. We study 5 different levels of progressive symmetric scatter plots having a common correlation, $R = 0.7$ to observe any underlying pattern between level of progressive symmetry and correlation perception. In order to bring variation in asymmetric shape whilst keeping R constant, we vary b and r simultaneously.

The following steps generate data points for SCP having correlation R. Each R is obtained by altering values of DR. Table in Appendix C gives different value of r and b corresponding to $R = 0.7$.

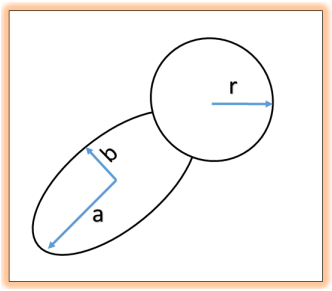


Figure E.2: Point cloud shape for task Progressive Symmetry

1. If $r > 0$, generate a data point, (x,y) such that x and y are random numbers between $+r$ & $-r$. If not go to step 5.

2. Check to see if(x,y) satisfies the equation of circle whose origin= (0,0) and radius= r : $x^2 + y^2 - r^2 < 1$.

3. If yes, translate the data point to shift it to the first quadrant at top right corner in the coordinate plane and add it to the set of data points for first circle else discard.

4. Repeat step 1 till the list has 40 data points in the set. Thus we have generated the circle of the cloud shape.

5. Again generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b$ and $-b$.

6. Check to see if (x,y) does not lie inside the circle AND also satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= b and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 < 0$.

7. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45°, just touching the circle and add it to the set of data points for ellipse else discard.

8. Repeat step 5 till the list has 40 data points (if r>0) or 80 data points (if $r = 0$) in the set.

9. Calculate correlation for all the 80 data points. If it matches the desired correlation $R = 0.7$, output the data set else go to step 1.

The same program is used to generate data set for all the trials of Task Distribution. Hence, we use the program to generate 5 SCPs each for $R = 0.7$ by varying r and b.

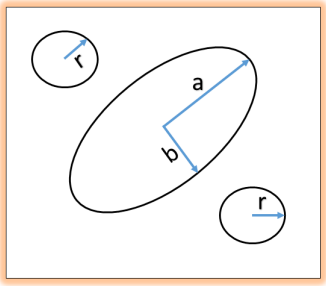


Figure E.3: Point cloud shape for task Distribution

E.1.1.4 Task4_Distribution.cpp

The main aim of task Distribution is to observe how perception of participant's vary with a change in distribution of points on the scatter plot. All the trials of this task has a cloud shape of scatter plot that is composed of 2 circles of same radius on the top left and bottom right corner of the scatter plot respectively and an ellipse in between at an angle of 45°. Figure E.3 shows one such scatter plot. Variation in distribution is achieved by altering the ratio of number of data points in the circles to the ellipse. In all variations, both the circles have equal number of data points. Hence DR is of the form $N_1 : N_2 : N_1$.

The following steps generate data points for SCP having correlation R. Each R obtained by altering values of DR. Table in Appendix C gives different value of a, r and b corresponding to R obtained from varying DR.

1. If $N_1 \neq 0$, generate a data point, (x,y) such that x and y are random numbers between $+r$ & $-r$, else go to step 6.

2. Check to see if(x,y) satisfies the equation of circle whose origin= (0,0) and radius= r : $x^2 + y^2 - r^2 < 1$.

3. If yes, translate the data point to shift it to the first quadrant in the coordinate plane in the top left corner and add it to the set of data points for first circle else discard.

4. Repeat step 1 till the list has N_1 data points in the set. Thus we have generated first circle of the cloud shape.

5. Again, generate a data point, (x,y) such that x and y are random numbers between $+r$ & $-r$ and follow the same steps done previously for the first circle to generate another set of N_1 data points with the exception that translation is done to shift data point to the bottom left corner of the quadrant.

6. Next generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b$ and $-b$.

7. Check to see if (x,y) satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= b and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 < 0$.

8. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45° and add it to the set of data points else discard.

9. Repeat step 6 till the list has N_2 (80 - 2*N_1) data points in the set.

10. Calculate correlation for all the 80 ($N_1 + N_2 + N_1$) data points. If it matches the desired correlation R, output the data set else go to step 1.

The same program is used to generate data set for all the trials of Task Distribution. Hence, we use the program to generate 6 SCPs each for $R = 0.25, -0.1, -0.3, -0.5, -0.7$ and -0.9.

E.1.1.5 Task5_Density.cpp

The main aim of task Density is to observe how participant's perception vary with a change in number of data points used to plot scatter plots. We perform this task using elliptical cloud shape and taking 5 different values of DP as reference points- $40, 60, 80, 100$ and 120. For each of these values, we observe perception at 3 positive correlation values- high (R=0.7), low (R=0.3) and neutral (R=0.5).

The following steps generate data points for SCP having correlation R and density DP and Table in Appendix C gives corresponding value of a and b.

1. Generate a data point, (x,y) such that x is a random number between $+a$ & $-a$ and y is a random number between $+b$ and $-b$.

2. Check to see if (x,y) satisfies the equation of ellipse whose origin= (0,0), semi-minor axis= b and semi-major axis= a : $\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 < 0$

3. If yes, rotate and translate the data point to shift it to the first quadrant in the coordinate plane at an angle of 45° and add it to the set of data points else discard.

4. Repeat step 1 till we list has DP data points in the set.

5. Calculate correlation for DP data points. If it matches the desired correlation R, output the data set else go to step 1.

E.1.1.6 Task6_Weber.cpp

We use the same code as that of **Task1_JND.cpp** to generate stimuli for this task. The cloud shape is elliptical and data points equal to 80. We generate stimuli with R= $0, 0.8, -0.1, -0.2, -0.5, -0.8$ and -0.9. Table in Appendix C shows the values of a and b corresponding to these R values. The data-set generation rules are same as that for task JND

All the data points generated using these C++ programs are then used to draw scatter plots using MS Excel. Before plotting, all the data set points are scaled upwards to lie between the range 1 and 100. The next section describes in detail the process of scatter plot generation and its subsequent storage for the software.

E.1.2 Scatter plot Generation

Once the data set has been rendered using the *.cpp files*, we use MS Excel to draw corresponding scatter plots after scaling the data points to lie in the range 1 to 100. Next, we copy the SCPs from MS Excel and '*paste special*' them as *Enhanced Metafiles (EMF)* picture in MS PowerPoint slides . We set the custom slide size to be 10 inch in width and 10 inch in height and lock the aspect ratio so that all the SCPs are identical in dimensions.

The entire file with all the slides is stored as a Portable Network Graphics (PNG) file, instead of PowerPoint Presentation (PPT) file. Doing so, creates a folder with all the images extracted from the slides, directly transported to it and saved as individual PNG files.

We then use Hyper Text Markup Language (HTML) to display the visualization stimuli i.e. the scatter plots by placing this folder in the same directory as the software program. The PNG files are then used to present the scatter plots in the experiment.

Bibliography

Afifi, A., May, S. and Clark, V. A. (2011). Practical multivariate analysis. CRC Press.

Alexiadis, M., Dokopoulos, P. and Sahsamanoglou, H. (1999). Wind speed and power forecasting based on spatial correlation models. *Energy Conversion, IEEE Transactions on*, **14**, 836–842. ISSN 0885-8969.

Altman, D. G. (1990). Practical statistics for medical research. CRC press.

Aschengrau, A. and Seage, G. (2008). Essentials of epidemiology in public health. Jones & Bartlett Learning.

Becker, R. A. and Cleveland, W. S. (1987). Brushing Scatterplots. *Technometrics*, **29**, pp. 127–142. ISSN 00401706.

Bizo, L. A., Chu, J. Y., Sanabria, F. and Killeen, P. R. (2006). The failure of Weber's law in time perception and production. *Behavioural Processes*, **71**, 201–210.

Bobko, P. and Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology*, **32**, 313–325.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, **6**, 151–155.

Cleveland, W. S., Diaconis, P. and McGill, R. (1982a). Variables on scatterplots look more highly correlated when the scales are increased. Technical report, DTIC Document.

Cleveland, W. S., Diaconis, P. and McGill, R. (1982b). Variables on scatterplots look more highly correlated when the scales are increased. Technical report, DTIC Document.

Cleveland, W. S. and McGill, R. (1984a). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, **79**, 531–554.

Cleveland, W. S. and McGill, R. (1984b). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, **79**, 531–554.

Cornsweet, T. (2012). Visual Perception. Elsevier Science. ISBN 9780323148214.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, **3**, 286–300.

Ding, Z., Granger, C. W. and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83 – 106. ISSN 0927-5398.

Doherty, M. E., Anderson, R. B., Angott, A. M. and Klopfer, D. S. (2007). The perception of scatterplots. *Perception & psychophysics*, **69**, 1261–1272.

Elmqvist, N., Dragicevic, P. and Fekete, J.-D. (2008). Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics, IEEE Transactions on*, **14**, 1539–1148.

Enns, J. T. and Di Lollo, V. (2000). WhatâĂŹs new in visual masking? *Trends in cognitive sciences*, **4**, 345–352.

Fraenkel, J. R., Wallen, N. E. and Hyun, H. H. (1993). How to design and evaluate research in education, volume 7. McGraw-Hill New York.

Friendly, M. and Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, **41**, 103–130.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, **45**, 135–145.

Halberda, J. (????). What is a Weber fraction.

Harrison, L., Yang, F., Franconeri, S. and Chang, R. (2014). Ranking visualizations of correlation using weber's law. *Visualization and Computer Graphics, IEEE Transactions on*, **20**, 1943–1952.

Jepsen, P., Johnsen, S. P., Gillman, M. and Sørensen, H. T. (2004). Interpretation of observational studies. *Heart*, **90**, 956–960.

Kanjanabose, R. (2014). An Empirical Study on Parallel Coordinates and. Ph.D. thesis, University of Oxford, UK.

Kolaczyk, E. D. and Csárdi, G. (2014). Statistical analysis of network data with R, volume 65. Springer.

Konarski, R. (2005). Judgments of correlation from scatterplots with contaminated distributions.

Kosslyn, S. M. (1985). Graphics and human information processing: a review of five books. *Journal of the American Statistical Association*, **80**, 499–512.

Lauer, T. W. and Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology*, **8**, 235–244.

Lew, M. J. (2012). Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *British journal of pharmacology*, **166**, 1559–1567.

Lewandowsky, S. and Spence, I. (1989). The perception of statistical graphs. *Sociological Methods & Research*, **18**, 200–242.

Li, J., Martens, J.-B. and Van Wijk, J. J. (2010). Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, **9**, 13–30.

Ligges, U. and Mächler, M. (2002). Scatterplot3d-an r package for visualizing multivariate data. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Loh, W.-Y. (1987). Does the Correlation Coefficient Really Measure the Degree of Clustering Around a Line? *Journal of Educational and Behavioral Statistics*, **12**, 235–239.

Maheshchandra, J. P. (2012). Long memory property in return and volatility: Evidence from the Indian stock markets. *Asian Journal of Finance & Accounting*, **4**, 218–230.

Meyer, J. and Shinar, D. (1992). Estimating correlations from scatterplots. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **34**, 335–349.

Mowry, T. and Luk, C.-K. (1997). Predicting data cache misses in non-numeric applications through correlation profiling. In Microarchitecture, 1997. Proceedings., Thirtieth Annual IEEE/ACM International Symposium on. ISSN 1072-4451, 314–320.

Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal*, **24**, 69–71.

Pearson, K. (1904). Mathematical contributions to the theory of evolution, volume 13. Dulau and co.

Pourhoseingholi, M. A., Baghestani, A. R. and Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from bed to bench*, **5**, 79.

Rensink, R. A. and Baldridge, G. (2010). The perception of correlation in scatterplots. In Computer Graphics Forum, volume 29. Wiley Online Library, 1203–1210.

Rubin, A. (2012). Statistics for evidence-based practice and evaluation. Cengage Learning.

Rubin, A. D. (1975). Hypothesis formation and evaluation in medical diagnosis.

Salkind, N. J. (2010). Encyclopedia of research design, volume 1. Sage.

Smeets, J. B. and Brenner, E. (2008). Grasping Weber's law. *Current Biology*, **18**, R1089–R1090.

Springett, K. and Campbell, J. (????). AN INTRODUCTORY GUIDE TO PUTTING RESEARCH INTO PRACTICE.

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, **9**.

Stern, M. K. and Johnson, J. H. (2010). Just noticeable difference. *Corsini Encyclopedia of Psychology*.

Stevens, S. (1975). Psychophysics. Transaction Publishers. ISBN 9781412832335.

Strahan, R. F. and Hansen, C. J. (1978). Underestimating correlation from scatterplots. *Applied Psychological Measurement*, **2**, 543–550.

Swallow, K. M. and Jiang, Y. V. (2013). Attentional load and attentional boost: A review of data and theory. *Frontiers in Psychology*, **4**. ISSN 1664-1078.

Tenenbaum, G. and Driscoll, M. (2005). Methods of Research in Sport Sciences: Quantitative and Qualitative Approaches. Meyer and Meyer Series. Meyer & Meyer Sport. ISBN 9781841261331.

Vergura, S., Acciani, G., Amoruso, V., Patrono, G. E. and Vacca, F. (2009). Descriptive and inferential statistics for supervising and monitoring the operation of pv plants. *Industrial Electronics, IEEE Transactions on*, **56**, 4456–4464.

Weber, E. H. (1978). EH Weber: The sense of touch. Academic Pr.

Yeshurun, Y. and Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, **396**, 72–75.

Zou, K. H., Tuncali, K. and Silverman, S. G. (2003). Correlation and Simple Linear Regression. *Radiology*, **227**, 617–628. PMID: 12773666.