chapter{Data collection}

Background on survey design

As explained by Wiley-Interscience [2004], a survey is a means of obtaining quantitative information regarding opinions and experiences of the respondents in order to explore the views of the target population as a whole. In this book, a survey is noted as a “systematic” method of collecting data, where the author states that the word “systematic” is deliberately used in order to separate surveys from other methods of information collection. “systematic” is defined by the Collins English Dictionary as something that *"is done according to a fixed plan, in a thorough and efficient way"* [Collins], and this reflects the manner in which surveys are created in accordance with a given system, where methods for distribution, implementation and analysis are defined under a pre-determined structure. The survey will be delivered to potential respondents in the target population, who will then be asked to complete a series of standardised questions, or questions for which the question ordering and wording is identical for every respondent, unless different formats are to be used to research purposes. It is once again discussed by Wiley-Interscience [2004] that standardised questioning was not always the norm; most interviewers would more likely have a list of objectives, and each interviewer would formulate and word questions based around these. It was discovered that question wording can have a drastic effect on respondents’ answers.

Whether or not the survey is ‘thorough’ and ‘efficient’ depends heavily on the survey structure and design. Designing an effective, systematic survey involves balancing efficiency with completeness, creating a survey that can obtain as much information as possible whilst not boring or fatiguing participants, which can lead to non-response and measurement errors due to participants skipping questions or selecting answers at random. A well-designed systematic survey has the capacity to yield large amounts of both qualitative and quantitative information regarding the research topic while minimising these errors.

There exist a variety of methods for delivering a survey, such as self-completed questionnaires and interviewer-administered interviews. Depending on the aims of the study, there will be advantages and disadvantages to each method. There may also be times when a combined approach is helpful in gathering the necessary information. The first method of surveying, a questionnaire, may consist of either physical paper forms that are mailed or handed out to people within the target population, or in an online format. As discussed by Brace [2004], this form of surveying constitutes a method of indirect communication between the respondent and researcher, in effect a non-verbal conversation in which the respondent is replying to the researcher’s questions. The non-face-to-face aspect of this method can be beneficial in terms of anonymity; an anonymous respondent is more likely to be honest in their answers than a respondent for whom the identity is known. As a result, an anonymous questionnaire can mitigate errors that may be caused by respondents fearing judgment of their answers. It is also possible to administer a large number of these questionnaires in a short period of time since they are self-administered, and thus constraints such as the number of interviewers or time taken to administer the survey has less effect on the amount of information obtained.

There are, however negatives to this questionnaire method. In his book, Brace discusses the way in which question wording must be very carefully thought about when using this method of indirect conversation, for reasons such as there being no way to correct participant misunderstanding of questions. Additionally, the fact that the researcher and participant never come into contact may allow the researcher to write questions without considering the human nature of the participants; it is easy to become absorbed in attempting to gather information and fall into forgetting that long-winded or complicated questions may bore or confuse respondents, leading to poorer quality responses. Similarly including too many questions in the questionnaire may lead to response errors for the same reasons. It is then crucial to be as clear and concise as possible in question wording, leaving little room for interpretation. This type of survey is also a very static medium; it does not allow for much expansion on participants' answers, with reasoning behind answers unknown unless specifically requested, which again could add to respondent fatigue and affect quality of response.

We can attempt to implement some dynamic discussion into a questionnaire in the form of 'open-ended questions', mentioned above as specifically requesting reasoning behind answers. A questionnaire is composed of two types of questions; closed-ended questions, for which the respondent selects their answer from a given set of potential responses, and open-ended questions, in which the participants are able to write their answers in a free-form format. Closed-ended questions are very good for obtaining quantitative data that may be easily categorised and counted, which is useful for gathering empirical evidence in order to form objective conclusions regarding the sample population.

Open-ended questions are generally used where more expansion may be required in addition to the closed-form answer, or if using a closed-form question would limit the answer range. The Leibniz Institute for the Social Sciences [Züll, 2016] provides guidance on open-ended questions, in which the occasions for using open-ended questions are outlined as:

- "knowledge measurement"; with with multiple choice, respondents would have a chance of guessing the correct answer, and thus this would be a sub-optimal way to measure raw knowledge

- "Unknown range of possible answers"; multiple choice may be limiting for certain questions, and may cause the researcher to miss important information

- "Avoidance of excessively long lists of response options"; if there is a known range of answers, but this range is very large, it may overwhelm respondents to see all of these as options

- "Avoidance of directive questions"; certain questions may have options based on the researcher's own opinions, and thus have the potential to direct the participant in a certain direction, and may not reflect the participants' true views. This links to "unknown range of answers" in that the researcher may incorrectly assume the potential range of answers and thus the given options may not cover the respondents' true opinions.

- "Cognitive pretesting", which covers instances such as ensuring the question was understood correctly.

To summarise, open-ended questions are useful when either there is not enough information to set a standardised range of potential responses or if more information is needed after a closed-ended response.

A method of surveying that is, by design, more dynamic is an interview. An interview may be structured, semi-structured or structured and each of these have a different set of features that distinguish them from one another. Structured interviews, as by the name, are rigid in nature and comprise of a vocal conversation in which the interviewer has a specific set of questions from which the discussion does not deviate. The slightly less rigid semi-structured interview is similar, but slight deviation from the plan is allowed in order to explore new avenues and ideas that might not be found with a structured interview, but the interviewer will still have a set of specific questions for which to obtain responses. For the most flexible of the three, the unstructured interview, the interviewer will tend to follow a loose plan of what they wish to explore rather than a strict question schedule, with the discussion led by the respondent's answers.

Phone calls and other forms of interview-based survey allow the interviewer to form a personal connection with the survey participant, which can be especially helpful for a company's image if the interviewer is particularly professional or charismatic. Additionally, while the interviewer will still be limited to asking the pre-set questions, the format of such a survey can be considered semi-structured and with much more room for interpretation. This can lend itself to gaining additional insights that may not have otherwise been gathered from a more closed-form paper or online survey. Additionally, the more open format can negate any error as a result of participants misinterpreting questions due to the interviewer's ability to immediately clarify on any misunderstandings. This type of survey also provides an instant response, which is beneficial if there is only a short time frame available in which to gather information.

However, there are also shortfalls to an interview-based survey method. For instance, although a charismatic interviewer can positively impact the image of whoever is conducting the survey, this could also lead to biases, such as the respondent answering in a way they feel will please the interviewer. Additionally, the image of the organisation could potentially be tainted if the interviewer appears rude or unprofessional, alongside potentially providing bias in the opposite direction. As well as this, telephone surveys are likely to be interpreted as a telemarketing scheme, and thus potentially have a negative impact on the number of willing respondents. The reduced anonymity of this type of survey may also create bias in the way of participants avoiding making statements that could be deemed socially unacceptable, or that they feel they may be judged for, and therefore may not provide answers accurate to their true line of thought.

The UK Household Longitudinal Study [lon] is an ongoing study and an example of implementation of a combined use of the above mentioned surveying methods. Initially, in 'wave

1’ of the study, a sample of 40,000 households in the UK were selected to be surveyed on a yearly basis. The survey involves all members of each selected household, overall comprising of around 100,000 individuals, and asks them a range of questions regarding areas such as family life, income, employment and health. The study consists of a self-administered youth paper questionnaire given to respondents ages 10-15, and an interview for those aged 16 and up. This split in age demographic allows some questions to be omitted from the youth survey, such as those about income and employment, and some to be added such as about pocket money habits and ‘future intentions’, as the website states. Giving the youth respondents a paper questionnaire may help obtain more useful or relevant answers, as the respondent may be more comfortable with this than being interviewed by an adult. The youth questionnaire is also shorter, which could perhaps just be a result of many questions not being relevant to this demographic, or it could be a conscious decision, but either way this with help to ensure the young respondent doesn’t lose interest and potentially incur bias in their answers due to either rushing to finish the survey or not paying attention. The adult survey also includes a section specific to 16-21 year olds. The surveys contain a standardised set of core questions asked each year alongside a set asked every other year. The reasoning behind this is given to be that this study has a very large scope, asking about many aspects of each respondents’ life, and so it becomes inefficient and counterproductive to include all questions every year since, as mentioned previously, the longer a survey is, the more likely a respondent is to get bored or mentally fatigued. The fact that the adult survey is administered in an interview also means that there may be limits on the amount of time the survey can take, as interviewers may have to get through a certain number of respondents in a day, additionally to the interviewer potentially also becoming fatigued. If the interviewer is fatigued, their tone and how they hold themselves may change, and potentially cause a subconscious bias in how the respondent answers the questions.

Specific goals of survey tool for this study

While visualisations can be a very useful tool for understanding data, they also have the potential to be highly misleading. This section of the study will explore how modifying certain aesthetic features of visualisations can impact perception and interpretation of data, and how these modifications can be exploited in order to mislead the observer. Misleading visualisations may be created in an effort to deliberately influence the viewers’ perceptions, or accidentally as a result of poor practice and knowledge surrounding data visualisation. In either case, visualisations have the ability to communicate different messages and stories depending on how they present the data to the observer. There is a large amount of research and literature surrounding this topic, both in terms of providing frameworks for good visualisation practice as well as looking into how various techniques are used to deceive viewers. Results from some of these papers will be replicated, as well as used to form hypotheses which this survey will investigate.

A large amount of the literature exploring misleading tactics in data visualisation focuses mainly on bar plots and line plots for categorical and time series data, and so this is what the survey will focus on. The specific aim of the survey is to test whether altering y-axis scaling, bar width, bar grouping method and colouring will have an impact on single data value interpretation and subjective interpretation of differences in data values.

Survey Design

The survey design will be inspired by a series of papers, all of which investigate how different aesthetic and design choices have the potential to mislead the observer or alter perception.

The 2020 paper “The Deceptive Potential of Common Design Tactics Used in Data Visualizations” [Lauer and O’Brien, 2020], as the title suggests, explores how using different design tactics may mislead the person seeing the visualisation. Similarly to “An Empirical Study of Data Visualisation”, the Claire and O’Brian paper uses a survey to explore how deceptive visualisation techniques can be employed as well as their impact on perception of the data. The survey discussed in this paper presents the participant with four plots; a bar plot, a line plot, a pie chart and a bubble plot. Additionally to changing aesthetic features of the plots themselves, the study investigates the use of exaggerated, leading titles, for example one control plot has the title " Home Sales Show Increase From 2015 - 2016“, which is altered to”Huge Increase in Home Sales From 2015 – 2016¡‘. The control plots consist of using a y-axis scaling beginning at 0 for the bar and line plots, a standard pie chart, and a bubble plot with proportionally sized bubbles, all alongside the non-exaggerated titles. The altered plots involve truncating the y-scale for the bar and line plots, making the pie chart in 3D, and arbitrarily altering the sizes of the bubbles on the bubble plot. The altered plots are referred to as the”deceptive" plots. The survey used sets of plots as crossed between deceptive aesthetics and deceptive titles; two had control aesthetics, one with the control title and one for the exaggerated title, and two had deceptive aesthetics with one having the exaggerated titling.

With regard to truncated axes, Claire and O’Brian asked participants to subjectively judge the difference between two data points using a 6 point scale ranging from “a little” to “a lot”. For both the bar plot and line plot it was found the the use of a truncated scale increases the perceived difference between the data points. The use of a truncated scale is also discussed by Yang et al. [2021], whereby 5 empirical studies were performed in order to assess the effect of altering the scale in this way. The first of the 5 studies once again assessed how large the difference between data points is perceived to be in the truncated plot as compared to a control, again using a subjective scale from “Not at all different” to “Extremely different” on a 7 point scale. This scale differed, however, in the way that a midpoint label of “Moderately different” was provided. The 7 point scale may be preferable to the 6 point scale as the 7 point has a defined midpoint at 4, whereas the 6 point does not. This study once again concludes that the differences in data points tended to be perceived as larger than for the control plot. Alongside these studies, a 2014 blog post [Parikh, 2014] discusses axis truncation and its effect on perceived data point difference for bar plots alongside other aesthetic features. The first example shows how truncating the y-axis of a bar plot can over-exaggerate differences in the heights of the bars, perhaps leading to incorrect observations regarding comparisons of values within the data.

The paper Hlawatsch et al. [2013] performs a similar study, but instead investigates the use of ‘stack-scale’, or ‘stacked’ bar charts and logarithmic scaling. The aim of the study was to explore whether stack-scale bar charts are an effective way to visualise large value data, which is less relevant to our study since we have relatively low-valued data compared to the paper, but nevertheless provides a framework for exploring the use of logarithmic scaling and stacked bars in a respondent study. Participants were shown three plots; a control with a linear scale, a bar plot using a stack-scale, and one with logarithmic scaling. The questions asked determined how the different scaling affected accuracy in reading individual values, interpreting differences in values and determining which time-step exhibits the largest difference in values. Motulsky [2009] additionally discusses the use of a logarithmic axis in bar plots, explaining how it is impossible for a zero value to be displayed on this axis, and thus the bar start points are arbitrary and produce an inaccurate representation of the bar height with relation to the true value. To quote the paper, *"Don’t create bar graphs using a logarithmic axis if your goal is to honestly show the data"*. We see that the logarithmic scale makes the percieved difference appear smaller than in the control.

Following this, questions included in part 1 the survey will focus on gauging whether altering the y-scale to be truncated or logarithmic has an effect on user perception of difference in data point values, for both bar and line plots. The respondents will be asked to gauge both individual values and differences in values, with the former providing an open answer box in which the may type their answer to allow for maximum freedom and obtain their true observation, unimpeded by the bias of having a specific set of numbers to pick from when their true observation may lie outside this range. The question for gauging difference perception follows Lauer and O’Brien [2020] and Yang et al. [2021] in using a numbered scale with numbers representing a range from not much difference up to a large difference. The Yang et al. [2021] method of a 7-point scale was employed here. From these papers, it is hypothesised that the truncated scale will cause respondents to overestimate differences between data values, and the logarithmic scale will be hypothesised to result in underestimation.

As well as scaling, another aspect of visualisation design that could potentially mislead the observer is bar width and aspect ratios. When adding a visualisation into a publication, re-sizing the visualisation to fit a specific gap may include altering the aspect ratio, in turn affecting the length to width ratio of the bars in a bar plot. As explored by Steven Few in a 2016 article for the *’Visual Business Intelligence Newsletter’* [Few, 2016], altering this ratio can affect viewer perception in the way of a narrower and taller image distorting bars to appear longer, and vice versa, meaning that perceived differences between bar heights may be affected. Part 2 of the survey was based around investigating this idea, alongside how the reading of exact values is affected.

Additionally, stacked bar charts will be investigated, showing a comparison between using the stacking method as opposed to a grouped bar plot. An article from the University of Stuttgart [Huynh, 2017] gives an overview of may types of bar chart, including stacked and grouped bars. The author remarks that grouped bar charts may make the comparison of bars in the same category more difficult, while the stacked bar chart sacrifices ease of comparison of values in the bars for increased spacial efficiency. A 2018 work from the journal of *'Visual Informatics'* [Indratmo et al., 2018] also provides a discussion on the use of various forms of stacked and grouped bar charts and their efficacy. The paper notes how a classical stacked bar chart can be useful for overall comparisons as the height of the bar represents the value of the item, with the different attributes depicted as a segmentation of this single bar into different colours. When discussing grouped bar charts it is mentioned that stacked bar charts may be less useful when performing attribute comparisons, in other words comparisons between different categories on the same bar, as a result of the bar segments being non-aligned. This results in comparison taking the form of length judgment as opposed to position judgment. Cleveland and McGill in their 1984 article in the *'Journal of the American Statistical Association'* [Cleveland and McGill, 1984] discuss how judgments based on length are likely to be less accurate than those based on position. A grouped bar chart is a way to allow for easy comparison between individual categories, but is discussed to be less effective in overall comparison. Based on this research, part 3 of the survey will include questions with the objective of testing standard stacked against grouped bar charts, alongside questions relating to the colour palettes used in depicting the different groups. We aim to test which colour palette is preferred in terms of aesthetics as well as ease of interpretation and reading.

The last two parts of the survey, noted henceforth as 'Sales - part 1' and 'Sales - part 2', explore the different y-axis scalings with respect to line plots, but for these, as opposed to the bar plots, the default was a truncated axis. The three plots investigated will consist of line plots relating to time series data for two fictitious companies. One will display each of the two lines on separate plots with the default axis, one will show both on the same plot with the default axis, and finally one with both on the same plot but with a zeroed axis. It is hypothesised that a difference in value for two time points will be perceived as smaller fopr the zeroed axis, and larger for the separated plots.

As discussed in Peytchev and Peytcheva [2017], too long a survey can result in higher measurement error due to factors such as waning interest or mental fatigue of respondents, resulting in careless responding and non-response. This is also further explored in Brower [2018], whereby a study is carried out to determine causes of careless responding, and specifically looks at questionnaire length and participant disinterest. The study performed in this work provides evidence that longer survey length can have a detrimental affect on careless responding; a long survey may make participants more likely to respond carelessly, and this must be considered when designing an effective and efficient survey. An additional conclusion states that participant interest in the survey content could have an effect, but also that evidence is less supported for this claim. There is significant enough evidence, however, to say that this should also be considered when designing the survey.

The Peytchev and Peytcheva [2017] paper explains that a 'split survey' design, where each respondent is only asked to answer a selection of questions from the whole set, is effective in reducing error while gathering large amount of information, however this will not be employed here. The reasoning for this is that there will already be a set of 12 different surveys being sent, and creating further splits could potentially lead to much too small sample sizes and thus inconclusive results. Additionally to this, the paper investigates how placement of questions in the survey can affect responses, concluding that questions asked later in the survey are more susceptible to bias, which tracks with the conclusion of survey length being a cause of careless responding; the longer a participant is taking a survey for, the more likely they are to start being careless with responding.

Due to this, the survey was designed to last in the range of approximately 15-20 minutes, as suggested in Revilla and Ochoa [2017]. One paper [Crawford et al., 2001] explores the pecieved burden of a survey on the participant, and performs a study whereby respondents were assigned a questionnaire, but given one of two different time estimates, for which the true length of the survey lay between. It was found that more people started the survey with the lower estimated completion time, but more also dropped out. However, the time at which respondents dropped out did not significantly differ in the two groups. In order to obtain maximum response, it is wise to as accurately as possible disclose the true survey length, and even slightly over-estimate in the disclosure.

With regard to the interest factor, the survey was designed with engaging respondents. The topic of the majority of the survey was chosen to be data relating to the television show *American Ninja Warrior*, as this could be subjectively viewed as a ‘more interesting’ topic than seemingly meaningless numbers. The survey was administered to a test subject, who commented that they found this topic interesting, with the additional comment that perhaps some pictures of the Ninja Warrior obstacles would be nice, however was not employed. The survey also took this respondent about 20 minutes to complete.

Although the content of the surveys for this study is not likely to be controversial or highly personal, anonymity is still important as the participants could otherwise potentially feel pressure to give a ‘correct’ answer, given the mathematical nature of the questions. As mentioned prior, anonymity here means that this pressure is potentially reduced and thus the relevant measurement bias may be mitigated. Additionally to the more technical visualisation questions, respondents were asked a series of demographic questions such as age, degree subject (if applicable), and whether they are colourblind or have any disorders that my affect visual processing. Additionally, three Likert scaled questions relating to well they would rate their spatial, observational and numerical skills. The Yang et al. [2021] paper, which explores the truncation effect of barplots, looks at graph literacy and its relation to perception, and hypothesises that those undertaking quantitative subjects at PhD level would be less impacted by the truncation effect as compared to humanities PhD students. It was found that the truncation effect did impact both groups, but those in quantitative fields had their perception marginally less affected. Thus the degree subject question was included to explore if this has an effect here. In relation to the visual processing and colorblindness questions, these are again included to test whether they have any significant impact on perception, as it may be important to consider these factors when creating visualisations to ensure they are accessible to all, and the study will examine the potential impact of such disorders.

The set will consist of two groups if surveys, which will be identical up to the visualisation package used. Particularly, one group will contain visualisations made with R’s ggplot2, the next with matplotlib from Python. These surveys will be distributed to the general public by sharing links on social media platforms such as Facebook. The reasoning behind creating two separate surveys in different languages is to ascertain whether the language used influences the interpretation. Within the groups there are 6 surveys, with each altering the order of visualisations shown in part 1 to assess the perception of each plot type without reference or comparison to another, and the same with part 2. in Part 3, each of the 6 used one of 3 colour palettes as the main colour, and another as a comparitor to test which the preferred colour palette is and which respondents find easier to read and interpret.

Creating the Visualisations

See appendix for the code and figures of the visualisations. The R visualisations were created using R version 4.0.2 [R Core Team, 2017] using ggplot2 version 3.3.3 [Wickham, 2009]. The Python visualisations were made using Python version 3.7.4 [Van Rossum and Drake Jr, 1995] with pyplot from matplotlib version 3.3.3 [Hunter, 2007].

The Data

The visualisations for the survey were created with inspiration from the papers discussed above. The bar plots were created using a data set regarding the history of obstacles used over 10 seasons of *'American Ninja Warrior'* [LAESSIG]. Each row of the data represents a single instance of an obstacle being used, and each instance has variables as specified in the below table.

Variable Name	Explanation
season	Season in which instance occured
location	Location of use
round_stage	Stage of competition in which instance occured
obstacle_name	Name of the obstacle
obstacle_order	Order in which the obstacle was placed in the course

This data was manipulated in R to produce a data frame containing the count of the number of times each obstacle was used over the course of the whole ten seasons. For the stacked and grouped bar plots, a data frame was produced, once again in R, containing columns 'obstacle' and 'stage', where 'obstacle' is a vector containing the name of each obstacle repeated the number of times it was used, and 'stage' similarly contains the names of all the stages of the competition, with each repeated the number of times it appeared. For example, Salmon Ladder was used 41 times, and thus is also repeated this many times, and there are 41 entries in the 'stage' vector corresponding to this. For the python version, the frequency tables were created manually.

The data for the time series plots was taken from the data set `BJsales` in the base R package `datasets` [R Core Team, 2017]. This data consists of a single vector of values with 150 entries, where each entry corresponds to a measurement taken at some arbitrary time point. Four subsets were taken from this data such that a start index was selected, and then this entry and the 11 following consecutive entries were extracted. The vectors were put into a data frame with the time steps set as months, giving a year of sales data for four fictional companies. This again was used to manually create a data frame in Python. To select the starting index, several seeds were tested for random selection, and four seeds were selected that would create plots to best test the hypotheses.

The Bar Plots

As explained before, the bar plots for part 1 were made such that one uses the default axis scaling, one uses a truncated axis, and one uses a logarithmically-scaled axis. It is worth noting that in R attempting to truncate the bar plot itself does not work; the bar must start at the zero tick mark otherwise the bars do not show up. To get around this issue, the data itself was truncated before applying to a bar plot with the tick labels then altered to fit the truncation, using intervals of 10 as in the default plot. Python, on the other had, will perform the truncation without this issue and defaults to steps of 2.5, which could affect the reading of values. For the logarithmically scaled plots, R by default starts at 1 and uses a non-standard form notation with tick labels of 1, 3, 20, 30. Python does use standard form and has labels 0, 10^0 and 10^1, starting at zero. The Python scale starting at zero was before mentioned as potentially misrepresenting the data. The height gauging of the R plot could maybe be impacted by the scale starting at 1. The default for the Python control plot scaling was more granular than the R, with steps on 5 as opposed to 10. The control scales for both languages have a range [0, 40], and [20, 40] for the truncated plots. There were 4 bars corresponding to 4 of the most used obstacles, arranged in descending order.

The next part plays with the aspect ratio of the plots. In order to keep this accurate, the plots were saved within the code as opposed to saving from the viewing window. The default aspect ratio for the ggplot is 1/1 for height to width, and using pyplot.gca() and comparing to the default we see that the default for Python using this method is 0.1. For the ‘wide’ plot, the aspect ratios are halved to 0.5/1 and 0.05, respectively. For the narrow, the aspect ratios were doubled to 2/1 and 0.2. Note that the aspect ratios include the entire plotting area, including labels and titles. These plots contained 7 bars as opposed to the 4, but were still arranged in decending order.

The plots in the third part of the survey were the stacked and grouped plots. The three colour schemes were the package default, a greyscale, and the colurblind-friendly Viridis palette [Garnier, 2018]. The obstacles here were the same 4 as displayed in part 1, but with the added colours for the competition rounds. The default axis ratios here mean that the R plots appear taller in comparison to their width than the Python plots, due to the legends.

The Line Plots

The plots for part 1 of this show the false sales data in the form of time series line plots, where the x-axis displays the months and y-axis shows number of sales. In the R version, the x-axis displays the 12 months in words, whereas the x-axis of Python version numbers the months and plots them in intervals of 2 months. This was an unintentional error on the part of the designer, however could be used to draw conclusions regarding how the two systems differ; monthly ticks in words or bi-monthly numbers. The plots in sales- part 2 were created very similarly, just with two different start indices.

The Survey

This section will discuss the specific survey questions and explain the differences in plot ordering and colour schemes between survey versions. Google forms was chosen as the medium for delivering the survey, as it is a free service and provides easy way to send out survey links and automatically compiles responses in a Google sheet along with time stamps, which can be exported to csv for analysis. To randomly assign each participant a survey, a javascript code was created to link to a landing page, which redirected the participant randomly to one of the 12 surveys. As time progressed it was possible to see how many respondents were taking each survey, and it was possibly to alter the Javascript accordingly to ensure each survey had an approximately even number of respondents. The survey was set such that each page contained a single question with a set of related sub-questions and only the plots relevant to these sub-questions, to prevent participants scrolling through the survey and seeing other figures which may alter their perception. This can also be used to analyse the effect of seeing other plots on perception of the plots following.

Demographic Questions

As discussed, the questions below are used to assess whether these factors have an impact on graph literacy and graph perception.

- Please enter your age (Open)
- If you are a university student or past university graduate please specify your area of study. (Drop down box: Science, Technology, Engineering, Maths, Arts, Social Sciences, Humanities, Business, N/A, Other (please specify))
- How strongly do you agree with each of the following statements? (Linear scale with 1 - 5, 1=strongly disagree, 5=strongly agree)
- - I have good spatial awareness skills
- - I have good observational skills
- - I have good numerical skills
- Are you colourblind? (Checkbox: Yes, No, Prefer not to answer)
- Do you have any disorders that may affect visual processing? (this could be a general visual processing disorder or dyslexia, dyscalculia, ADHD etc) ((Checkbox: Yes, No, Prefer not to answer))

American Ninja Warrior - Part 1

The questions regarding each of the three bar plots were as follows:

- Approximately many times would you say the ’Salmon Ladder’ was used? (Open)
- Approximately how much more than ’Log Grip’ would you say ’Salmon Ladder’ was used? (1-7 scale)
- Approximately how much more than ’Quintuple Steps’ would you say ’Salmon Ladder’ was used? (1-7 scale)
- In your opinion, approximately how many times would you say ’Log Grip’ was used, as a percentage of the number of times ’Salmon Ladder’ was used? (Open)

Here, the two questions with the difference rating scale are used to assess whether having the bars next to each other vs on opposite ends of the plot has an effect on the difference in rating when comparing the responses for each of the plots.

The table below shows all the permutations of the three plot types, and which questionnaire version they appear in.

	Q1	Q2	Q3
V1	Control	Log	Truncated
V2	Control	Truncated	Log
V3	Log	Control	Truncated
V4	Log	Truncated	Control
V5	Truncated	Control	Log
V6	Truncated	Log	Control

The table shows that, for example, in version 1, the control plot was shown in question 1, the log-scaled in question 2 and the truncated in question 3.

American Ninja Warrior - Part 2

The questions regarding each of the three bar plots were as follows:

- How large would you say the difference between ’Jumping spider’ and ’Salmon Ladder’ is? (1-7 scale)
- How large would you say the difference between ’Log Grip’ and ’Floating Steps’ is? (1-7 scale)
- How many times would you say ’Floating Steps’ were used? (Open)

Similar to part 1, the below table gives all permutations of the three plot types.

	Q1	Q2	Q3
V1	Default	Narrow	Wide
V2	Default	Wide	Narrow
V3	Narrow	Default	Wide
V4	Narrow	Wide	Default
V5	Wide	Default	Narrow
V6	Wide	Narrow	Default

Questions regarding comparisons between the plots were then administered as follows, while showing respondents all of the three plots on a single page.

- Which of the three bar charts do you find most aesthetically pleasing? (Multiple choice with options "A", "B" or "C")
- Which bar chart do you feel is easiest to read and interpret? (Multiple choice with options "A", "B" or "C")
- Which bar chart do you find hardest to read and interpret? (Multiple choice with options "A", "B" or "C")

American Ninja Warrior - Part 3

This part explored the differences in perception for stacked and grouped bar charts, alongside colour preferences. This part had 4 questions, with the first two asking about the stacked and grouped bar plots, with either the stacked first or grouped first.

The first two sub-questions are given below.

- How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds? (Open)
- How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds? (Open)

The next question is “Please select the statement you feel applies to the bar chart above.” and consists of a multiple choice answer with the following options:

- 'Log Grip' was used MORE in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used Less in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) rounds and Qualifying (Regional/City) rounds.")

This is followed by another mulitple choice question, given as Which obstacle do you think was used MORE in Finals (Regional/City) rounds, ‘Log Grip’ or ‘Floating Steps’?, with the following options:

- ’Log Grip’
- ’Floating Steps’
- They were used the same amount of times

After answering these questions for both plot types, the respondents were shown both on the same page and asked to select which of the two they found easier to read and interpret, and were then shown the stacked bar plot in two different colour palettes; the one used for the questions so far and a comparitor, with the questions below.

For the stacked vs grouped comparison:

- Which bar chart do you feel is easiest to read and interpret? (Multiple choice with options "A", "B", "C")

For the colours comparison:

- Which colour scheme do you find most aesthetically pleasing? (Multiple choice with options "A", "B", "C")
- Do you feel that one of the colour schemes makes it easier to read and interpret the data than the other? If so, please select which one. (Multiple choice with options "No", "Yes, A is easier", "Yes, B is easier")

For this part, survey versions 1, 2 and 4 showed the stacked bars first, followed by the grouped, and versions 3, 5 and 6 displayed the grouped first. It is shown in the below table which colour schemes were used in each survey.

Version	Main colours	Comparitor
V1	Viridis	Default
V2	Default	Viridis
V3	Default	Greyscale
V4	Greyscale	Default
V5	Viridis	Greyscale
V6	Greyscale	Viridis

Sales - Part 1

The respondents then moved onto part 1 of the sales section of the survey, in which they are asked to once again give subjective opinions regarding the y-axis scaling, but this time relating to time series line plots.

Once again, the same set of questions is asked for each plot which consist of, firstly, a two-row multiple choice grid, with each row relating to one of the companies. Respondents were asked the question “How much would you say sales of each company increased between January and December?” and were to give a response on the 7-point scale.

The ordering of the plots for each version number are given below.

	Q1	Q2	Q3
V1	Separated	Truncated	Zeroed
V2	Separated	Zeroed	Truncated
V3	Truncated	Separated	Zeroed
V4	Truncated	Zeroed	Separated
V5	Zeroed	Separated	Truncated
V6	Zeroed	Truncated	Separated

The second question was “How large would you say the drop in sales between April and July of Company A is?”, which once again was rated based on the 7-point scale.

Sales - Part 2

The final part of the survey showed zeroed and truncated plots once again, for two different fictitious companies, this time with the intention of gaining an overall view. For each of the two, each respondent was asked a single 7-point scale rating question; “Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?”.

Bibliography

Understanding society - the uk household longitudinal study. URL https://www.understandingsociety.ac.uk/.

Ian Brace. *Questionnaire Design: How To Plan, Structure And Write Survey Material for Effective Market Research.* 01 2004.

Cheyna Katherine Brower. "too long and too boring: The effects of survey length and interest on careless responding". 2018. URL https://corescholar.libraries.wright.edu/etd_all/1918.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. ISSN 01621459. URL http://www.jstor.org/stable/2288400.

Collins. Systematic. In *Collins.com dictionary.* URL https://www.collinsdictionary.com/dictionary/english/systematic.

Scott D. Crawford, Mick P. Couper, and Mark J. Lamias. Web surveys: Perceptions of burden. *Social Science Computer Review*, 19(2):146–162, 2001. doi: 10.1177/089443930101900202. URL https://doi.org/10.1177/089443930101900202.

Steven Few. Bar widths and the spaces in between, 2016. URL http://perceptualedge.com/library.php.

Simon Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2018. URL https://CRAN.R-project.org/package=viridis. R package version 0.5.1.

M. Hlawatsch, F. Sadlo, M. Burch, and D. Weiskopf. Scale-stack bar charts. *Computer Graphics Forum*, 32(3pt2):181–190, 2013. doi: https://doi.org/10.1111/cgf.12105. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12105.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Hai Dang Huynh. Two-dimensional bar charts. 2017. doi: http://dx.doi.org/10.18419/opus-9496. URL https://elib.uni-stuttgart.de/handle/11682/9513.

Indratmo, Lee Howorko, Joyce Maria Boedianto, and Ben Daniel. The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons. *Visual Informatics*, 2(3):155–165, 2018. ISSN 2468-502X. doi: https://doi.org/10.1016/j.visinf.2018.09.002. URL https://www.sciencedirect.com/science/article/pii/S2468502X18300287.

MATT LAESSIG. Anw obstacle history. URL https://data.world/ninja/anw-obstacle-history.

Claire Lauer and Shaun O'Brien. The deceptive potential of common design tactics used in data visualizations. In *Proceedings of the 38th ACM International Conference on Design of Communication*, SIGDOC '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375252. doi: 10.1145/3380851.3416762. URL https://doi.org/10.1145/3380851.3416762.

Harvey J. Motulsky. The use and abuse of logarithmic axes. Online, GraphPad Software, inc., 2009. URL https://web.archive.org/web/20101123050530/http://graphpad.com/faq/file/1487logaxes.pdf.

Ravi Parikh. How to lie with data visualization, Apr 2014. URL https://heap.io/blog/data-stories/how-to-lie-with-data-visualization.

Andy Peytchev and Emilia Peytcheva. Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, 11(4): 361–368, Dec. 2017. doi: 10.18148/srm/2017.v11i4.7145. URL https://ojs.ub.uni-konstanz.de/srm/article/view/7145.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Melanie Revilla and Carlos Ochoa. Ideal and maximum length for a web survey. *International Journal of Market Research*, 59(5):557–565, 2017. doi: 10.2501/IJMR-2017-039. URL https://doi.org/10.2501/IJMR-2017-039.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Wiley-Interscience. *Survey Methodology (Wiley Series in Survey Methodology)*. 01 2004.

Brenda W. Yang, Camila Vargas Restrepo, Matthew L. Stanley, and Elizabeth J. Marsh. Truncating bar graphs persistently misleads viewers. *Journal of Applied Research in Memory and Cognition*, 2021. ISSN 2211-3681. doi: https://doi.org/10.1016/j.jarmac.2020.10.002. URL https://www.sciencedirect.com/science/article/pii/S2211368120300978.

C. Züll. Open-ended questions. *GESIS Survey Guidelines.*, 2016. doi: 10.15465/gesis-sg_en_002.