This chapter will discuss basic univariate analysis of the survey results, including summary statistics and univariate testing for the whole population as well as the subsetting for the programming language used and degree type. Additionally, subsets will be created considering only the first plot shown for each question, drawing comparisons between responses for these plots themselves without influence of the others. The analysis will be performed in R version R version 4.0.2 [@R].

In terms of testing, Shapiro-Wilk tests will be applied with the `shapiro.test()` function to gauge whether the data sets can be considered normally distributed and thus whether parametric T-Tests are suitable for either one-sample or paired comparisons, for the Shapiro-Wilk test, the alternative hypothesis is that the data is not normally distributed. Failing he normality condition, a symmetry test will be administered via the `symmetry.test()` function from the package `lawstat` [@lawstat], and providing there is insufficient evidence to reject the null hypothesis that the data is symmetric, a Mann-Whitney-Wilcoxon (MWW) test will be used. If there is sufficient evidence that data proves neither symmetric nor normally distributed, sign tests will be applied. MWW will also be used for two sample testing where perhaps a sign test would be most appropriate, but cannot be used as the samples are of different sizes.

The sample sizes are 70, 38 and 32 for the whole population, R subgroup and Python subgroup, respectively before removing NA of invalid values. The sample means and medians will be notated as $\bar{x}$ and $\tilde{x}$, respectively.

American Ninja Warrior - Part 1

This part of the survey assess the effect of truncated and logarithmic scaling on bar plots perception and interpretation.

The final question in part 1 of the survey, *'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?'* will not be considered as it is similar to the previous questions, and responses ranged in form, between percentages and decimals, and it can not just be assumed that all the decimals can be converted to percentages; for example a value of 0.5 could be the decimal value for 50%, or the respondent could have meant this as 0.5%.

Effect of Y-Axis Truncation

In general, truncating the y-axis had less of an effect than anticipated. In question 1, *"Approximately many times would you say the 'Salmon Ladder' was used?"*, for which the true value was 41, the distribution of responses for the truncated plot ($\bar{x} = 41.35$) as compared to that of the control plot responses ($\bar{x} = 41.21$) shows a small difference, with the mean perceived value of the bar being slightly higher for the truncated plot. The median for both of these is 41, showing that

both distributions are centered around the true value of 41. The control and truncated plots have contextually fairly small variances of 0.752 and 0.753 respectively, depicting both that there is limited variation in the responses and most of the observations lie fairly close to the respective means. The variances are also quite similar, showing that the distributions appear fairly similar, as emphasised by observing the below density plot.

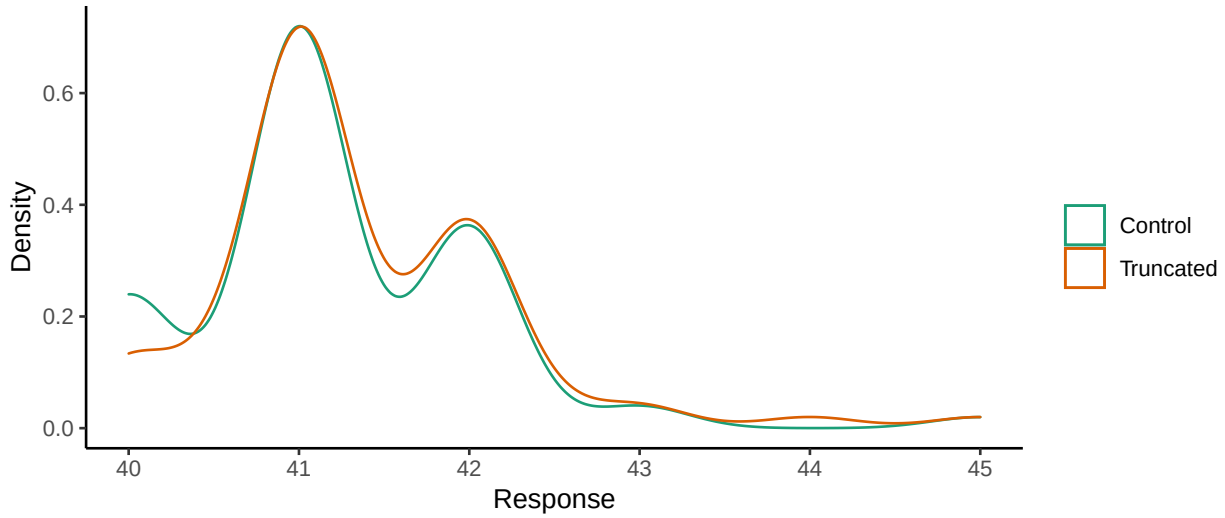


Figure 1: Density plot showing distributions of responses regarding the control and truncated plots for the question 1

Performing a dependent-samples sign test comparing these two sets of responses confirms that there is no significant difference ($p = 0.1877$) in the response distributions. However, the one sample sign tests show that there is not sufficient evidence to suggest the control plot responses differ from the true value of 41 ($p = 0.1214$), but there is evidence to accept the hypothesis that the truncated plot responses differ from the true value ($p = 0.0026$). This shows that, while there is insufficient evidence from sign testing to suggest a statistically difference in the responses for the two plots, the location of the truncated plot responses may be slightly further from the true value than the control, and it is confirmed by a one sided sign test with an alternative hypothesis that the true median of truncated responses is greater than 41 ($p = 0.0002$). This gives evidence that the truncated plot results in a slight overestimation in reading of the bar height as compared to the true value of 41. Note that in the responses for the control plot for question 1, there was a response of "41/41", which was taken to be 41.5.

In question 2, *'Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was was used?'*, the set of responses for the truncated plot ($\bar{x} = 5.87$, $\tilde{x} = 6$) is considered significantly different by a dependent-samples sign test from the control plot responses ($\bar{x} = 5.36$, $\tilde{x} = 5$). By eye, the average values do not seem too different between the two plot types, although the p-value

of the sign test ($p = 0.00019$) shows that there is in fact a statistically significant difference. The perceived difference for the truncated plot being rated higher on average than for the control plot provides evidence to accept the hypothesis that using a truncated scale can cause differences in bar height to appear larger, once again this is confirmed by a one-sided sign test ($p = 9.554e-05$), with the alternative hypothesis that the true median of truncated responses is greater than that of the control responses.

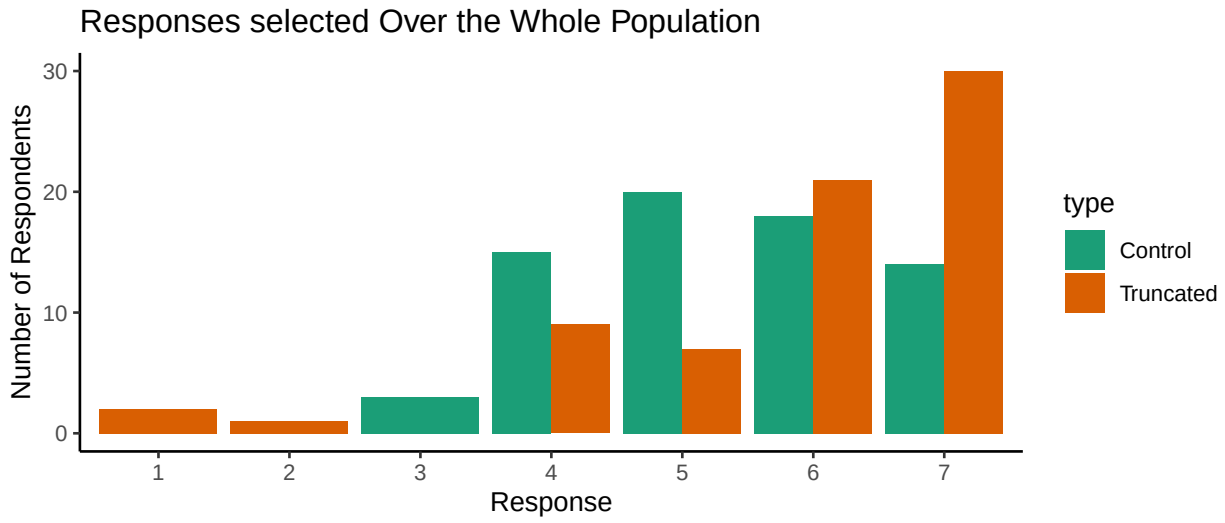
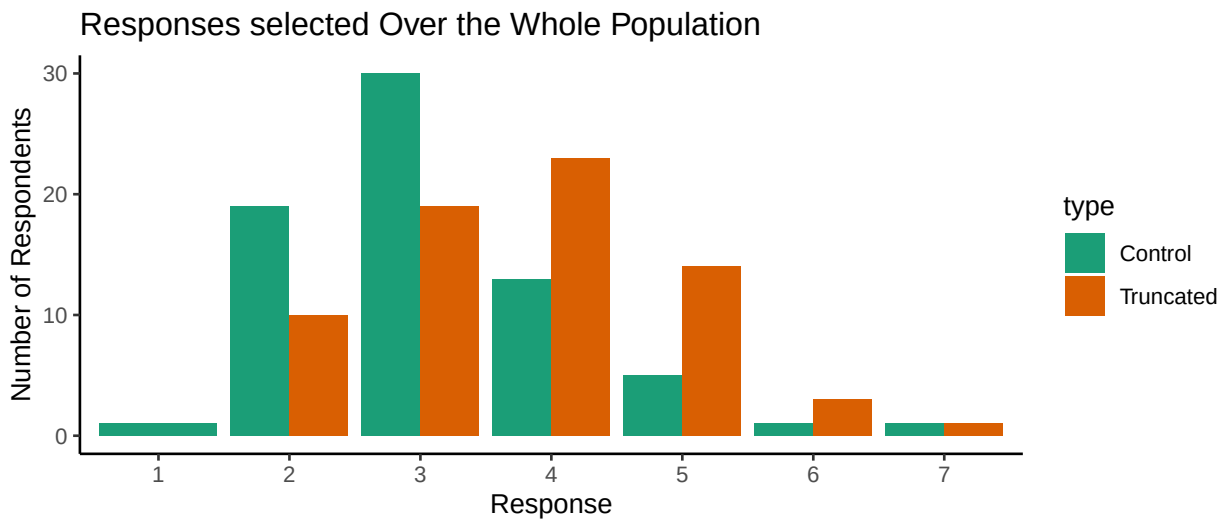


Figure 2: Bar plot showing distributions of responses regarding the control and truncated plots for question 2

The spread for the truncated and control plot responses are slightly skewed to the right, depicting that the subjective view on the difference between the bar heights was that it was in general on the larger side. Looking at the bar heights, for the responses of 4 and 5 the control plot bars are higher, and vice versa for the truncated plot response bars. This again emphasises the evidence to support the hypothesis that truncation leads to larger perceived difference.

Question 3 of part 1, *'Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?'*, asks a similar question to question 2, but asks respondents to judge the difference for bars on opposite ends of the plot as opposed to next to each. Again, the by eye comparison shows not a massive difference between distributions of responses for the control ($\bar{x} = 3.12$, $\tilde{x} = 3$) and truncated ($\bar{x} = 3.12$, $\tilde{x} = 3$) plots, although the sign test shows that the there is evidence to suggest that the truncated plot responses are in fact on average greater than for the control plot ($p = 4.624e-06$). figure[?] shows the distribution of responses.



The response distributions, conversely to question 2, now seem skewed more to the left. However there is a similarity in the way that for the lower ratings of 2 and 3, the control plot response bars dominate, and for the responses of 4 and 5 the opposite is true.

Overall, it seems that the use of truncation has a small but statistically significant effect on perception of height difference between bars, with respondents tending to judge the difference as slightly larger than for the control plot, although this effect is smaller than initially anticipated, and larger for bars that are further apart. In terms of reading values from bars, the truncation did not have a statistically significant effect when comparing the two distributions, however in one sample testing the truncated plot responses did differ significantly from the true value.

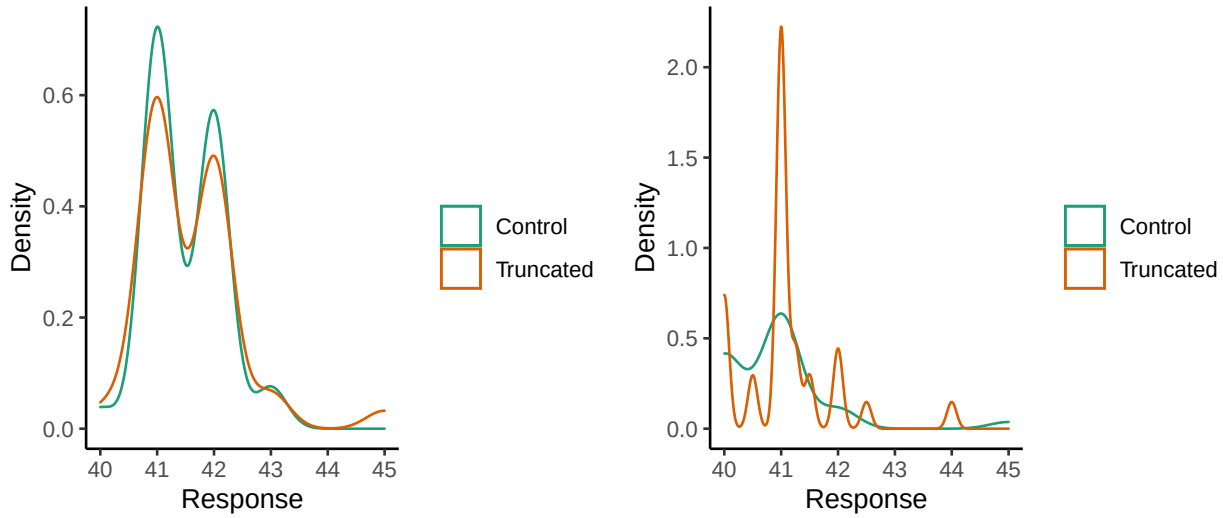
When considering the language subgroups, note that there is a discrepancy here between languages in terms of the axis tick breaks and labeling, with the R plot being incremented in steps of 10 for both the control and truncated plots and the Python being more granular in steps of 5 for the control and steps of 2.5 for the truncated.

Consider question 1. Comparing the two language subgroups for the truncated plot, the distributions for both the R ($\bar{x} = 41.56$, $\tilde{x} = 41$) and Python ($\bar{x} = 41.01$, $\tilde{x} = 41$) responses to question 1 appear similar in location to those of both each other and the whole population ($\bar{x} = 41.35$, $\tilde{x} = 41$).

Comparisons via MWW testing show that the responses related to the control plot differ statistically significantly between the two language cohorts ($p = 0.00012$), and similar for the truncated plot responses ($p = 0.02163$), where the tests were performed comparing first the R and Python responses for the control plot, and then for the truncated.

A sign test shows sufficient evidence that the R subgroup responses relating to the truncated plot differ from the true value ($p = 0.0004$), whereas there is insufficient evidence when applying a MWW test to the Python responses ($p = 0.718$). Similarly, the R subgroup's responses in relation to the

control plot statistically significantly differ from the true value ($p = 7.629e - 05$), but the Python subgroup's do not ($p = 0.1185$). This could potentially be a result of the less granulated R plot scaling, due to the reduced precision.



The distributions for the control and truncated plot responses for the R subgroup are fairly similar to the whole population, although the peaks for the logarithmic plot responses are marginally lower. The distribution of the truncated plots is unexpected from lokking at the numbers, and more 'chaotic'. This shows potentially more variation in the responses.

For question 2 it is similarly seen that the language used does not have a statistically significant impact on the response for the truncated plot, with means 5.500 and 5.187, and medians 6 and 5 respectively for R and Python for the control plot, and means 5.98 and 5.84 both with median 6 for the truncated. Comparative testing with MWW gives $p = 0.2199$ for the control plot and 0.9105 for the truncated. Thus, the scale granulation or any other differing aspect of the plots does not seem to have a significant effect. See figure[?] for the distributions.

For question 3 (see figure [?]), it is again seen that the responses in relation to the R version of truncated plot ($\bar{x} = 3.76$, $\tilde{x} = 4$) do not differ significantly to those related to the Python version ($\bar{x} = 3.78$, $\tilde{x} = 4$), with a two sample MWW p-value of 0.9708. Similarly the control plot, there is little difference between the R ($\bar{x} = 3.342$, $\tilde{x} = 3$) and the Python ($\bar{x} = 2.87$, $\tilde{x} = 3$) versions of the plot, again with am MWW p-value 0f 0.1465.

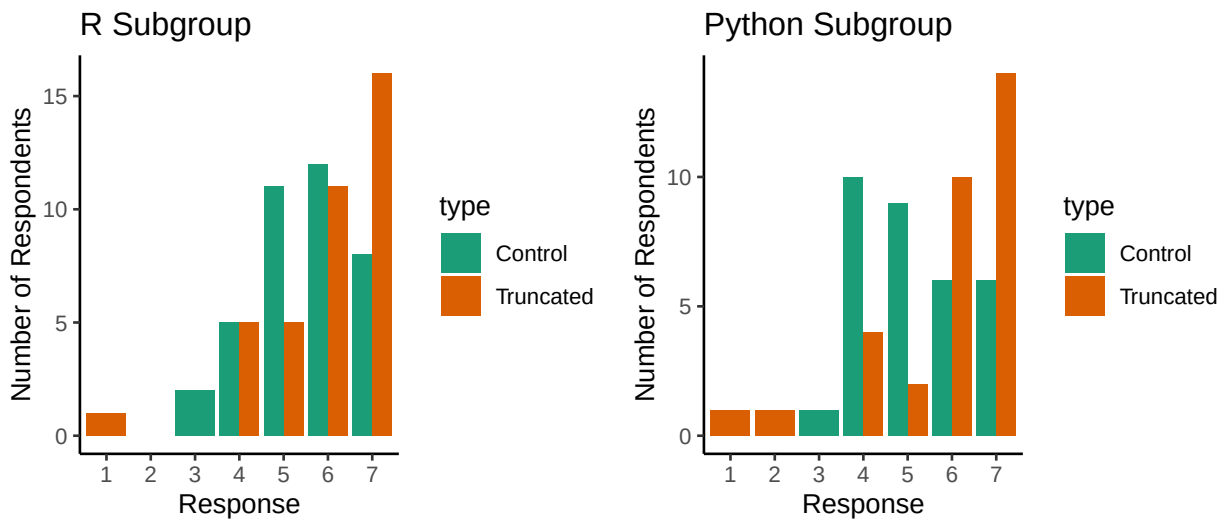
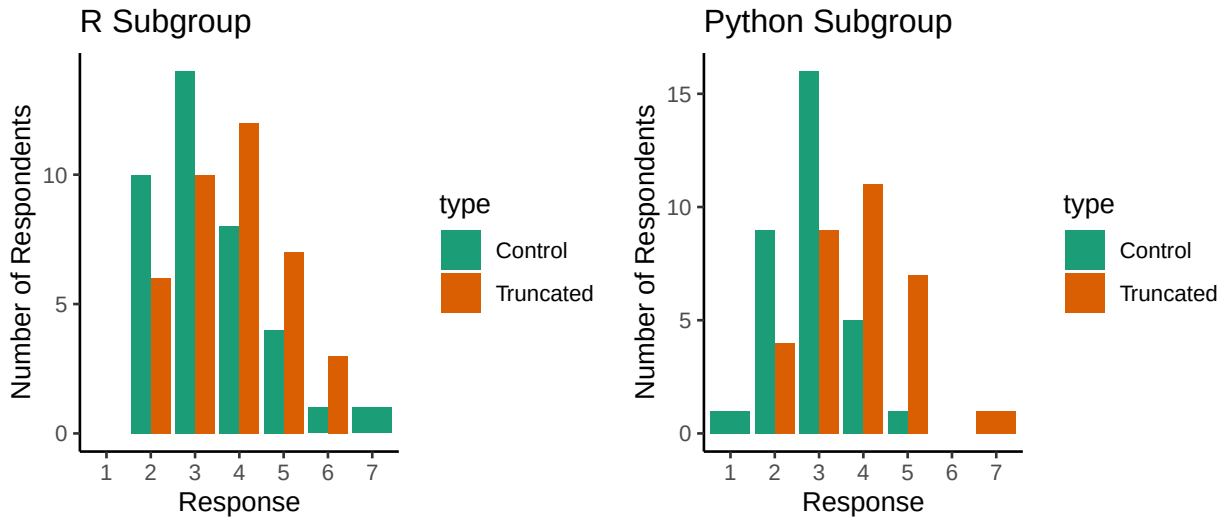


Figure 3: Bar plot showing distributions of responses regarding the control and truncated plots for question 2, for the R and Python subgroups



Figure[?] shows both distributions, with the R appearing more positively skewed and the python looking fairly symmetric for both plot types, which was also found when performing symmetry tests. For the Python it can also easily be seen that the bars for the truncated plot responses seems 'shifted' to the right slightly as compared to the control.

Now considering subsetting for the respondents that saw the truncated plot first out of the three. Note that 25 saw the control plot first and 23 saw the truncated plot first.

The distribution of responses for the truncated plot in question 1 shows a slightly higher mean (41.696) and median (41.25) than for the whole population, but a MWW test shows that the

difference is not significant ($p = 0.1379$). Similarly for questions 2 and 3, performing tests on the truncated plot for respondents who saw this first as compared to the truncated plot responses for the whole population result in p-values of 0.2614 and 0.3145, providing evidence that the plot order doesn't have much of an impact on perception for the truncated plot.

The conclusions appear to be consistent with results from the @YANG2021 paper, in which the researchers, similar to this survey, showed participants a series of control bar plots alongside those with a truncated axis, and concluded that the difference in values for the truncated axis were perceived to be larger than those of the control plots.

Effect of Logarithmic Scaling

Within the logarithmic responses, there were two invalid responses, given as 'Don't know' and 'Next to none.'. These will be considered as 'NA' responses and discounted from the quantitative analysis, however they do provide useful qualitative insights into how the respondents reacted to the plots, particularly as both were entered for the logarithmically scaled plot made in Python.

The mean of the responses for the logarithmically-scaled plot, on the other hand, was magnitudes higher than the true value at 1.493e+13, although with a median of 35; lower than the median response of the control and truncated plots responses. The high magnitude is the result of two answers of '10^15' and '10^9', both again for the python version of the plot.

The default logarithmic scaling in Python uses standard form notation, which perhaps the two participants who entered the high magnitude answers were less exposed to and not as familiar with. Looking at the degree subjects for these respondents, it is observed that they study Social Sciences and Psychology, respectively. This could add to the idea that they are less familiar with this notation as it is more commonly used in mathematical and physical science disciplines. One of the respondents also rated their numerical skills at 1/5, showing they feel that numerical skill is not their specialty. The other rated their numeric skills at 4/5, showing that even with a good self-perceived level of numerical skill, standard form could be considered misleading.

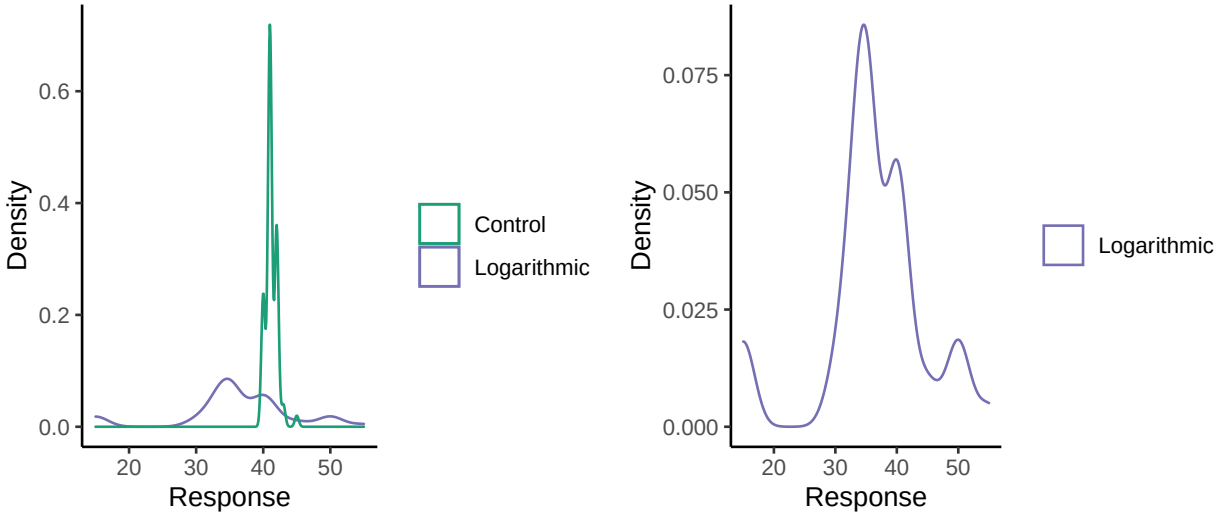
This should perhaps be considered when designing visualisations; the creator of the visualisations may find the logarithmic scale or standard form more effective in showing the data, but they should consider the target audience. Are the audience going to be familiar with this? If, for example, visualisations are being published in a paper targeted at academics in a subject likely to use such scalings often and understand them, this may be a good way to depict the data. However, using this in something such as an advertising campaign could mislead the public, causing them to either over or under estimate values. As previously discussed, however, this is often done deliberately in order to push the message the creator wishes to sell.

The variance in the responses for the logarithmic plot is also high, with value 1.492×10^{28}, showing

that a large amount of the observations differ from the very high mean, and considering this alongside the lower median may point towards many of the respondents either giving an accurate response or even underestimating. Furthering this point, the IQR for the logarithmic responses is the interval $[30, 40.5]$, which sits below the true value, displaying that over 50% of the observations in the total population actually underestimate the value.

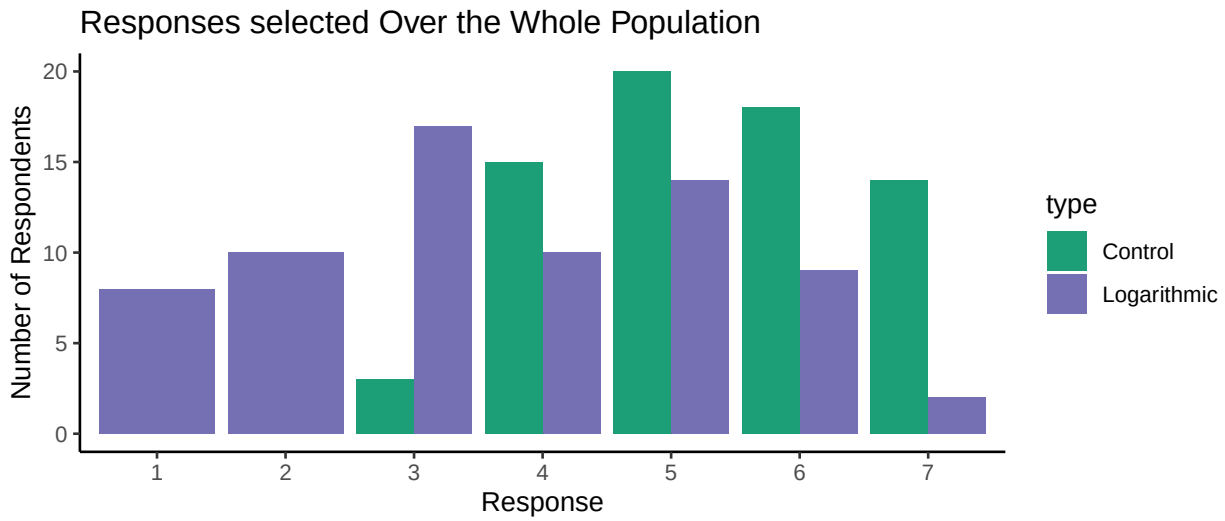
The distribution of responses in the R subgroup also shows on average a slight underestimation ($\bar{x} = 39.73$, $\tilde{x} = 35$) and, as expected, vast overestimation for the Python version ($\bar{x} = 39.73$, $\tilde{x} = 35$). This shows that, with a linearly notated logarithmic scale, the scale may cause underestimation, but this is counteracted by using a standard form notation.

It can be considered to follow the convention of values that have value outside the range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$, where $Q1$ and $Q3$ are the first and third quartiles, which here would be the range $[14.25, 60.75]$ and results in a sample size of 59. Consider now the response distribution for the logarithmically-scaled plot, after removing these responses, for which figure[?] gives the density plot. Both plots show the response distribution of the outlier-removed set of responses, with the providing a comparison with the distribution of responses relating to the control plot.



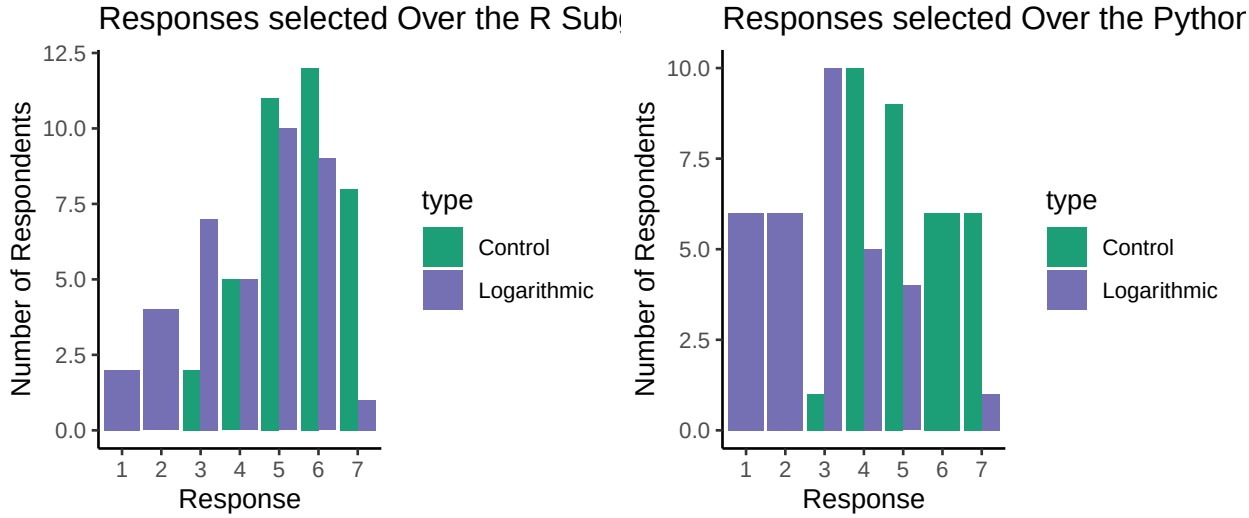
The Python default of standard form notation appears to have confused certain respondents, who are perhaps not as used to seeing this notation, and there was a large range in the responses along with one person not even entering a number, but rather stating that they "Don't know", and another stating they believed the value was "Next to none". The "Next to none" entry is subjective, but could potentially be be assumed as a value close to 0, once again maybe as a result of standard form being less well known to this respondent.

The distribution of responses for question 2 is displayed in figure[?].

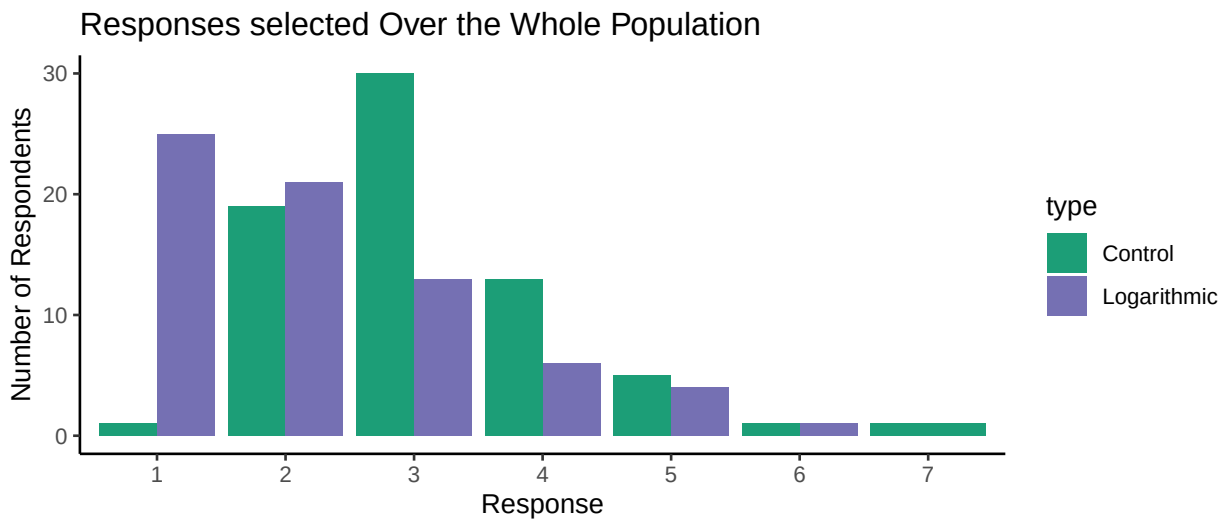


The spread of logarithmic plot values is fairly wide, with at least one response for each option, and the control is the same as stated before. The plot depicts how there is a wide spread of values, with some respondents having very different subjective views of the size of the difference to others. On average, the subjective perceived difference in bar heights was significantly lower for the logarithmic plot responses ($\bar{x} = 3.67$, $\tilde{x} = 3.5$) than for the control ($\bar{x} = 3.35$, $\tilde{x} = 5$). This is evidenced by a one-sided sign test with the alternative hypothesis that the logarithmic plot responses are on average lower than the control plot responses.

There is evidence to show that the difference between the R and Python versions of the logarithmic plot is significant ($p = 0.00096$, $\bar{x}_R = 4.263$, $\bar{x}_{Py} = 2.969$). The distributions for the two language subsets are shown in figure[?].

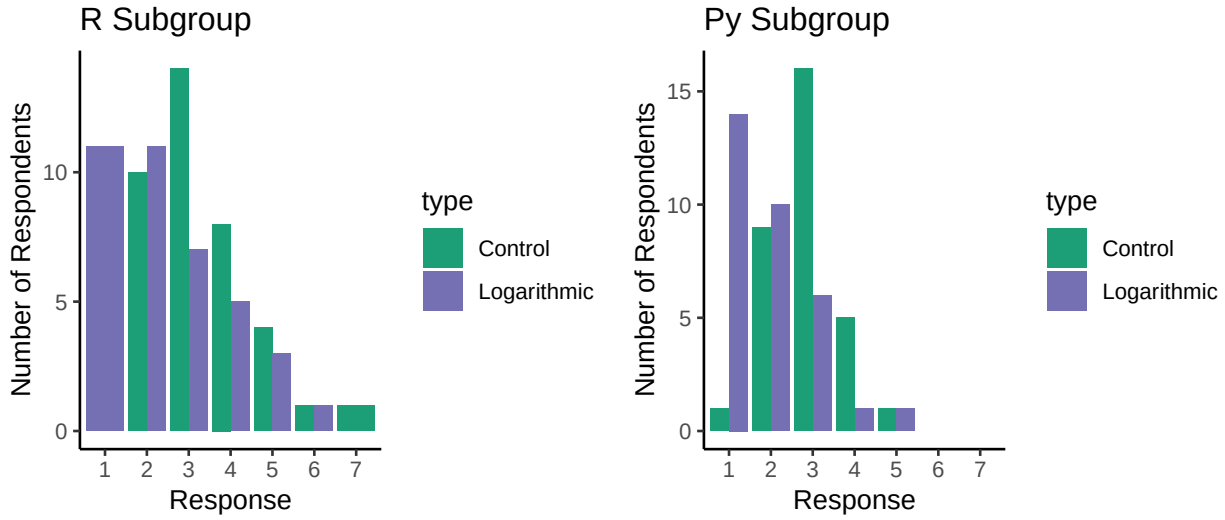


In regard to question 3, see again figure[?] for the plotted distributions.



The responses for the logarithmically scaled plot are skewed towards the lower end of the scale, similar to the control and truncated responses, and there does not appear to be much difference between distributions of the two populations. Looking at the numbers, however, the averages for the logarithmic plot ($\bar{x} = 2.22$, $\tilde{x} = 2$) seem lower than that of the control plot ($\bar{x} = 3.77$, $\tilde{x} = 4$). Indeed, a one sided MWW test comparing the logarithmic and control plot responses elicits a p-value of $1.317e - 06$, showing evidence that the logarithmic scale resulted in lower rating in difference of bar height.

Figure [?] shows the distributions for R and Python subgroups.



The distributions of the logarithmic plot responses for the R ($\bar{x} = 2.5$, $\tilde{x} = 2$) and Python ($\bar{x} = 1.9$, $\tilde{x} = 2$) subgroups appear fairly similar, with the same median albeit with the mean for the R subgroup being slightly higher. The plots to appear to show the R subgroup responses being slightly positively skewed and the Python responses more centered around 3. A two sample, one sided

MWW test provides sufficient evidence that the R responses appear in average greater than the Python ($p = 0.03689$).

Looking at the responses from the respondents who saw the logarithmic plot first of the three, the average responses from this group for question 1 ($\bar{x} = 40$, $\tilde{x} = 40$) were closer to the true value of 41 than for the whole population ($\bar{x} = 36.277$, $\tilde{x} = 35$), although the former still differs significantly from the true value ($p = 6.104e - 05$), and there is not significant evidence to state that the two distributions differ ($p = 0.1705$). Comparing the response statistics for the whole population and for those who saw the logarithmic plot first, the log first group perhaps show the bar height difference being perceived slightly higher than for the whole population ($\bar{x}_{overall} = 3.67$, $\bar{x}_{logfirst} = 4.13$), however a two-sample MWW test gives an insignificant p-value of 0.2614 when comparing them. Similarly, the difference between the responses for the whole population and for those who saw the logarithmic plot first for question 3 is also statistically insignificant, with means of 3.08 and 2.68 for and a p-value of 0.1889.

Differences Between Question 2 and 3 Responses

Now take $\bar{x}_{control} - \bar{x}_{truncated}$ and $\bar{x}_{control} - \bar{x}_{logarithmic}$ for each of questions 2 and 3, which is shown in figure[?].

Table 1: Table showing difference in the percieved difference for the logarithmic-scaled and truncated plots as compared to the control, for questions 2 and 3

	Con - Trnc	Con - Log
Q2	-0.5142857	1.685714
Q3	-0.6428571	0.900000

This again shows that the responses for the truncated plot were in general rated higher than the control plot responses, and also that the effect was more significant for the bars on opposite ends of the plot as compared to the bars next to each other. The opposite is true for the logarithmic plot responses; on average they were rated lower than the control plot, but this was greatly more significant for the bars next to each other, as opposed to the truncated plot. Figure[?] shows this visually.

On average, truncating the scale had a similar effect for both questions, albeit with slightly more effect for when comparing 'Salmon Ladder' with 'Quintuple Steps' as opposed to 'Log Grip'. For the logarithmically scaled plots, however, the re-scaling appears to have had a significantly greater effect when considering the bars directly next to each other, with respondents on average judging the difference in bar height to be greater by 1.68 on the 7-point scale, whereas this is 0.9 for the bars further apart. It can be concluded from this that truncating the scale had more of an impact

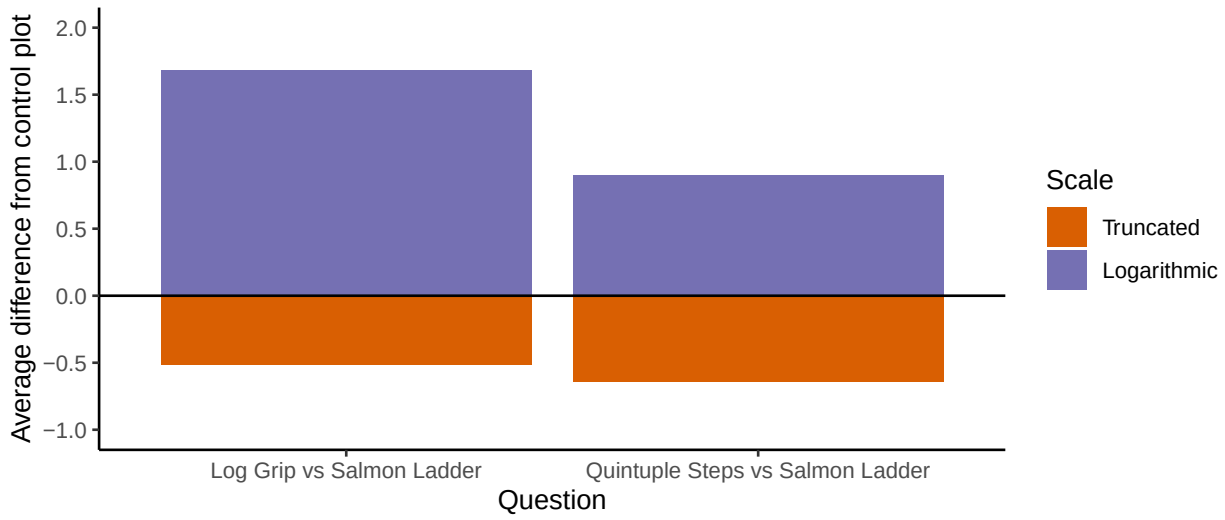


Figure 4: Bar plot giving a visual representation of the table

when bars were on opposite ends of the plot as opposed to next to each other, and the way round for the bars close to each other; the logarithmic scaling had more of an impact.

Ninja Warrior - Part 2

This part of the survey assessed whether different aspect ratios would have an impact on perception of bar height differences as well as reading of true values. This part will be analysed question by question.

Question 1 asked *'How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?'*. This question once again uses the 7-point scale to gain a subjective view on the degree to which respondents felt the heights between the two bars corresponding to 'Jumping Spider' and 'Salmon Ladder' differed for three bar plots of 7 obstacles, where 'Salmon Ladder' is furthest to the left, and 'Jumping Spider' furthest to the right.

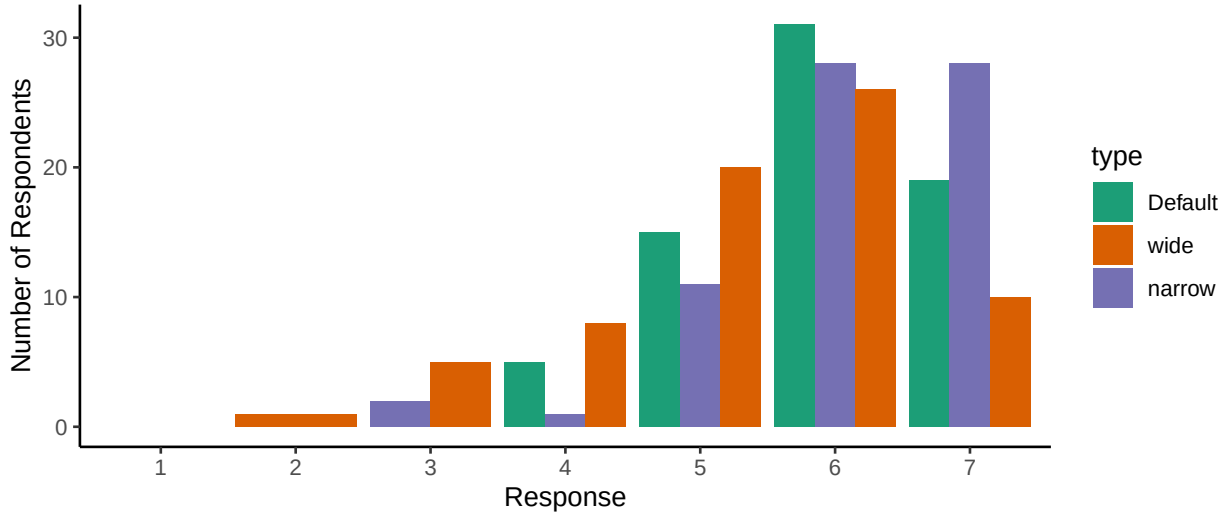
Looking at the means and medians here, it doesn't seem like there is that much of a difference in perception of the differences between the three aspect ratios, as displayed in table[?].

	Default	Narrow	Wide
Mean	5.914	6.129	5.357
Median	6.000	6.000	6.000

Note that 'narrow' is defined as the plot with the aspect ratio of smaller width to greater height, and vice versa for the 'wide' plot. The means show marginal differences, whereby the default plot mean is the middle-valued mean of the three, with the mean perceived difference for the wide plot being

slightly smaller than this and the mean perceived difference for the narrow plot is slightly larger. This result, although at first glance marginal, follows the hypothesis that the wide plot would cause differences to be perceived as smaller and narrow bars to cause differences to be perceived to be greater.

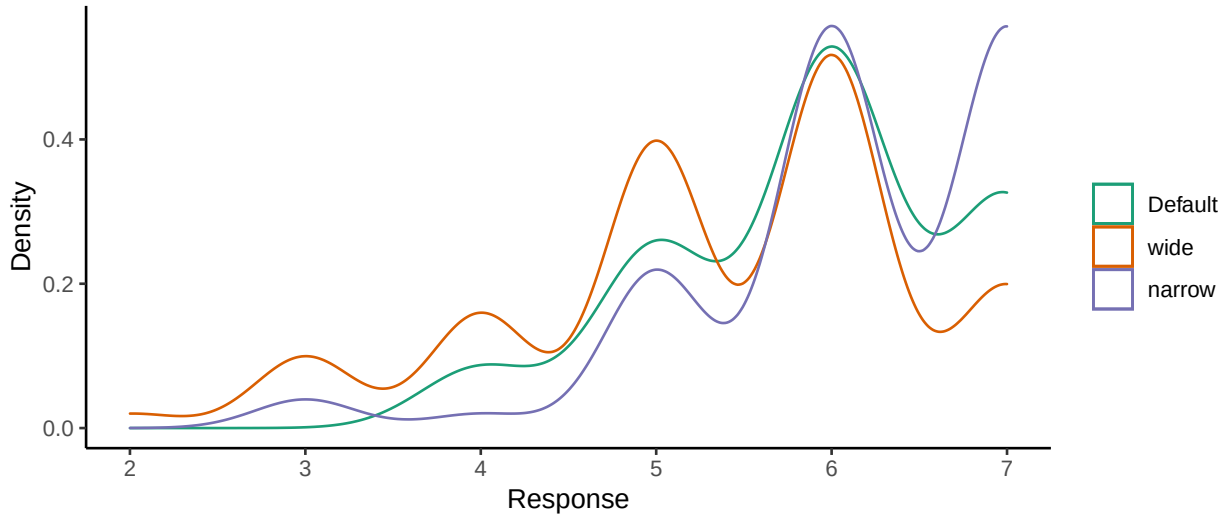
Now looking at figure[?], showing the three distributions. There isn't an immediately obvious difference in distributions, but on closer inspection it can be seen that the orange "Wide" bars dominate over the three for the range [2, 5], and the purple "Narrow" dominated for the response of 7, following the above analysis of summary statistics. There was a fairly strong consensus that in general that a rating of 6 was applicable to all three plots.




```
    Dependent-samples Sign-Test

data:  default and wide
S = 31, p-value = 6.457e-06
alternative hypothesis: true median difference is greater than 0
95 percent confidence interval:
   0 Inf
sample estimates:
median of x-y
            0

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
```

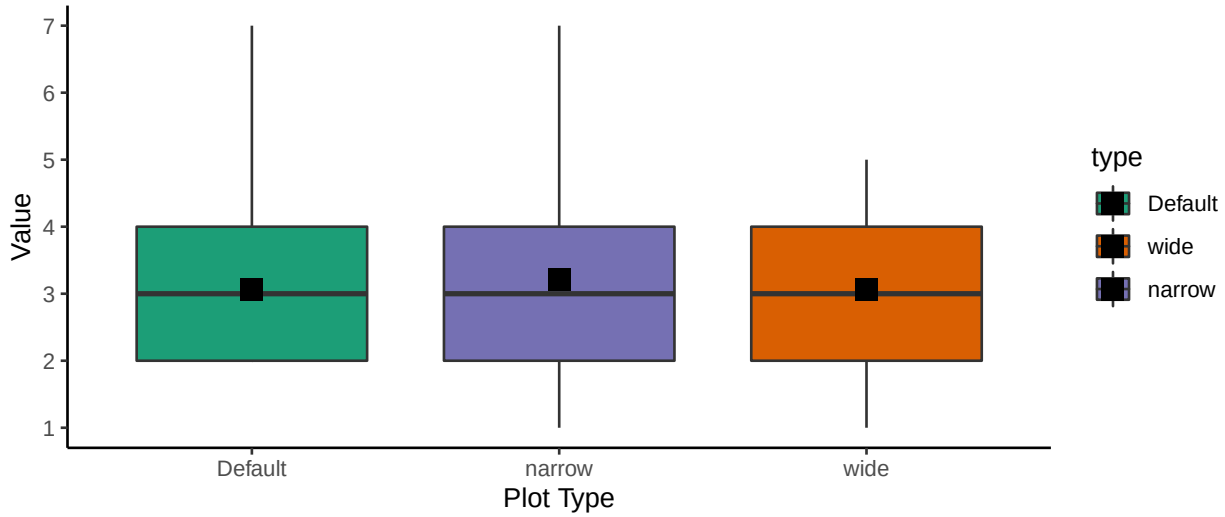
```
Lower Achieved CI     0.9402      0    Inf
Interpolated CI       0.9500      0    Inf
Upper Achieved CI     0.9639      0    Inf
```

Running a one-sided MWW test to compare the responses for default plot to the narrow plot, it is confirmed that there is evidence to suggest that using a 'narrow' aspect ratio causes the perceived difference to be greater ($p = 0.0468$). Then applying a one-sided sign test to compare the default to the wide plot, the perceived difference is shown to be smaller ($p = 6.457e - 06$).

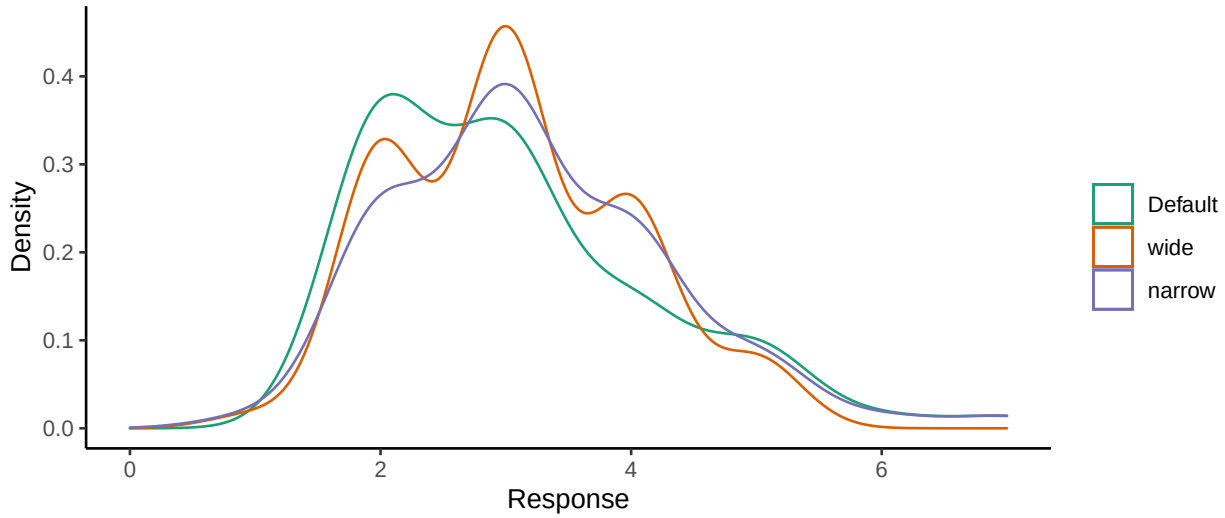


Question 2 then went on to ask *'How large would you say the difference between 'Log Grip' and 'Floating Steps' is?'*. Similar to part 1, there are two questions for gauging differences between bars, for which one asks about bars far away from each other, and one about bars next to each other. In the case of this section, the first question contained bars on opposite ends of the x-axis, and this question asks about two bars that sit adjacent to one another.

The analysis results here show that altering the axis ratio appears to have even less of an effect than in the first question, with the means of the responses for the default and wide plots being identical at 3.057, with the mean of the narrow plot responses only 0.157 greater at3.214. The median for all three is 3, and the IQRs are all $[2, 7]$. The variances, however, do differ from one another, with values 1.301, 0.866 and 1.214 for the default, wide and narrow bars, respectively. The distribution of values are shown in figure[?]. The results of two-sided MWW tests show that neither aspect ratio appears to have a significant effect on the rating of the perceived difference ($p = 0.2446$ and $p = 0.5688$).



At least 50/ of respondents placed the difference in the range $[2, 4]$ for all three plots, showing that they believed the difference was small to moderate, and this didn't change depending on the plot type, and thus for the bars further apart from each other, changing the aspect ratio does not appear to make much of a difference. The overall distributions are shown in the figure[?]



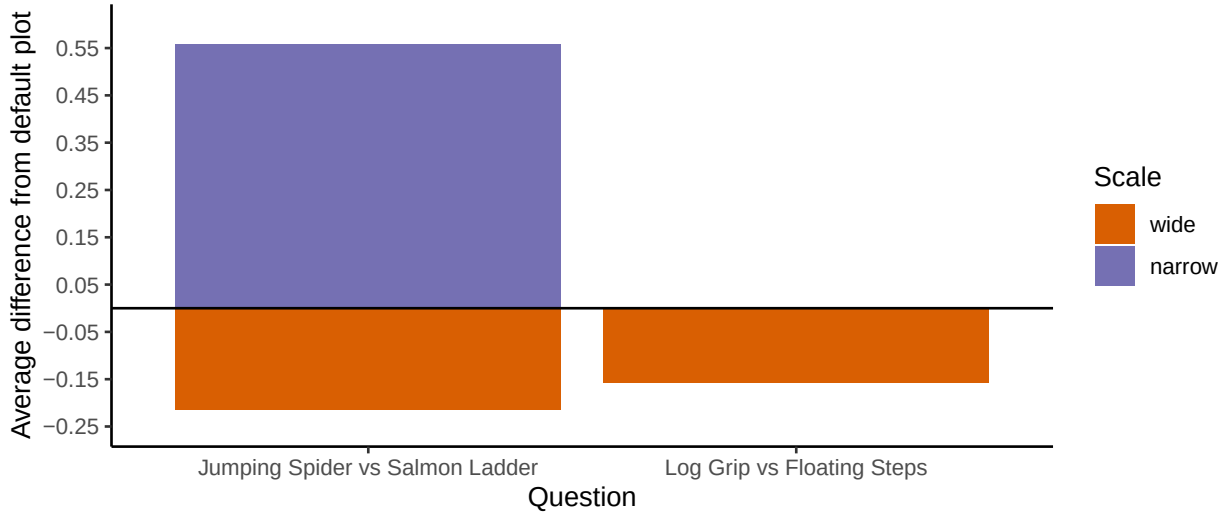
All three distributions are very similar, and almost appear to form bell curve shaped distributions, albeit with some irregularities and very slight negative skew.

As in part 1, the two height difference perception questions will be compared, calculating $\bar{x}_{default} - \bar{x}_{narrow}$ and $\bar{x}_{default} - \bar{x}_{wide}$.

As before, the table below gives a visual representation.

Table 2: Table showing difference in the percieved difference for plots with narrow and wide bars as compared to the default, for questions 1 and 2

	Def - Narrow	Def - Wide
Q1	-0.2142857	0.5571429
Q2	-0.1571429	0.0000000

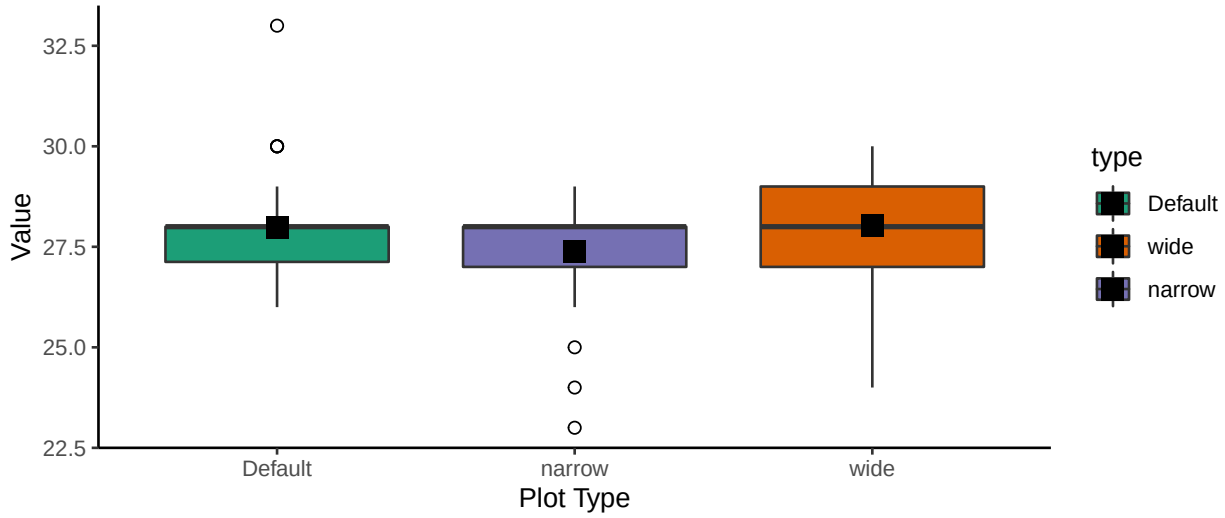


Both by eye comparisons of values and statistical testing show that the language used has negligible effect on the perceived difference, as does the order in which the plots were shown. See tables 51 - 61 in the appendix for more details.

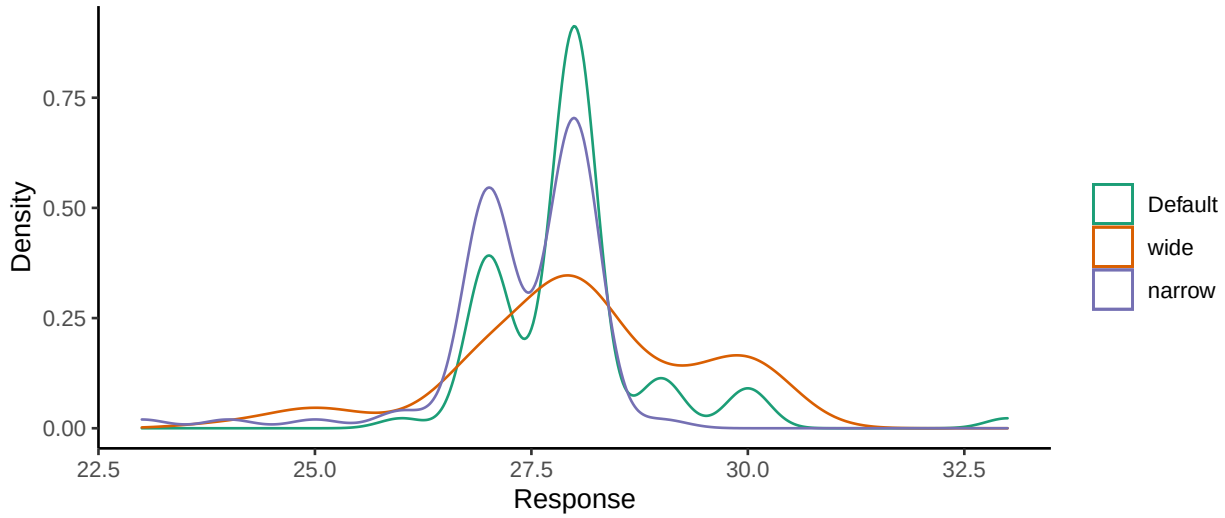
How many times would you say 'Floating Steps' were used?

This is again similar to question 1 of part 1, where participants were asked to state what they believed to be the height of the bar for 'Salmon Ladder', however this time the third bar from the axis is chosen. This is to ascertain whether the distance of the bar from the axis may have an effect alongside any potential perceived distortion of values. Note that the true value was 28.

The means of each of the three sets of responses were very close to the true value, at 27.97, 28.04 and 27.39, respectively for the default, wide and narrow, and the medians are exactly equal to the true value. Based on the means and medians it appears that, once again, altering the aspect ratio had minimal, if any, effect on interpretation of the data value. The value for the default plot also appears to be closer to the true value than the control plot in part 1, question 1.



Looking at the box plots, there are very small ranges in the values, signifying that there was a large consensus between respondents in terms of what they perceived the height to be. It can also be seen that there are three outliers below the box plot for the narrow plot responses, and two above for the default plot responses. There is very little overlap between the boxes, and it appears again that there altering the aspect ratio of the bar plot has little to no impact on reading the height of the bar. Additionally, there was less agreement between respondents for the wide plot than for the other two, although this doesn't seem to be too significant.



The distributions for the default and narrow plot responses are similar, both seeming to be fairly centred on the mean with a steep decrease in density on either side of the mean to shallow tails within the range $[25, 30]$. The responses for the wide plot appear to be more spread with lower density function values, with a slight negative skew.

After removing the outliers the medians have stayed the same, and the mean has obviously decreased

for the default and increased for the narrow, however, these means are all still fairly similar to each other and at a first glance prior to testing it again seems that changing the aspect ratio, at least to the degree tested here, is inconsequential to interpretation of the actual value. As expected as well, the variances for the outlier-removed sets have decreased.

However, statistical tests do actually show that while the default responses did not differ significantly from the true value of 28 ($p = 0.5667$), the responses for the narrow plot did ($p = 2.0955e - 09$), but the wide didn't ($p = 0.5067$).

Changing the language and plot order was once again inconsequential here.

Differences Between Question 1 and 2 Responses

The last set of questions in part 2 show respondents all three of the bar plots presented in this section and ask them to select which they find most aesthetically pleasing, and which they find easiest and hardest to interpret. Table[?] gives the number of respondents that selected each plot for each of the three questions.

	Default	Narrow	Wide
Most aesthetically pleasing?	37	14	18
Easiest to read and interpret?	36	15	19
Hardest to read and interpret?	20	20	30

For the first question, relating to how aesthetically pleasing respondents found each plot, just over half of the respondents chose the default aspect ratio as the most aesthetically pleasing, with 37 out of the 69 who responded selecting this.

Similarly, 37 out of the 70 that responded to the second question found the plot with the default aspect ratio easiest to read and interpret. Perhaps the people that preferred this aspect ratio aesthetically did so because they found it easiest to interpret. Investigating this, 27 respondents who chose the default for question 1 also chose this for question 2.

The plot judged hardest to read and interpret by the most respondents was the one with the wide bars, with 30 selecting this and 20 selecting each of the other two. While a significant number chose the default and narrow bars, the slightly higher amount selecting the plot with wide bars matches the previously stated hypothesis formulated from following the Stephen Few paper, which discusses that an ratio of greater width to length could suffer from perceptual imbalance. While this imbalance isn't seen in the numbers from the previous questions, the result here does give some indication that the aspect ratio producing wide bars may impact on ease of interpretation.

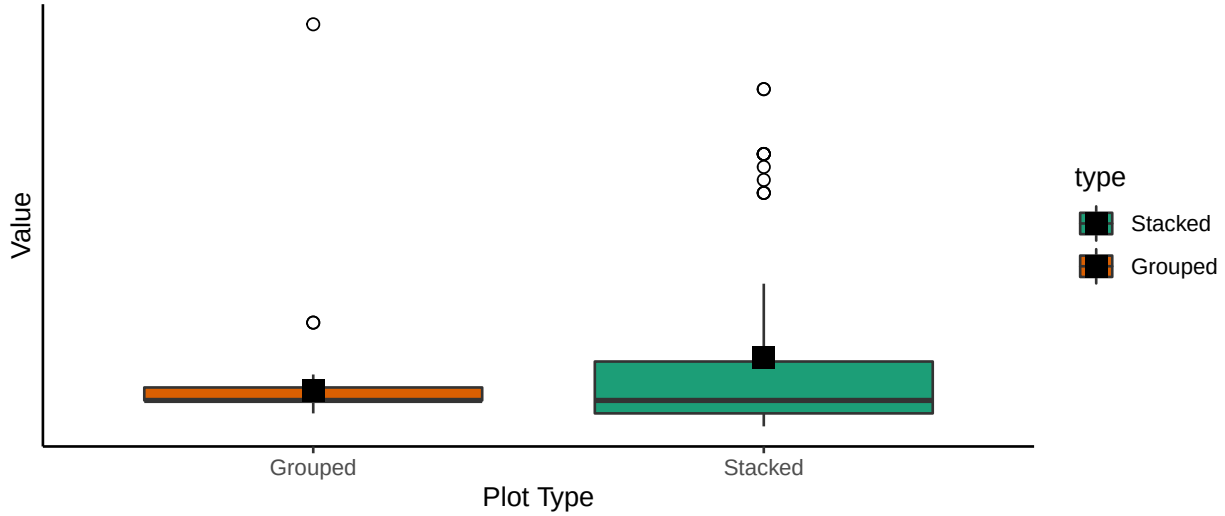
Ninja Warrior - Part 3

The third and final part of the questions about the American Ninja Warrior data discusses stacked bars and colour schemes. The questions asked in this part are used to decipher how data with multiple categories may be best represented in a bar plot. The plots presented use the same bars as in part 1, but this time the number of times each obstacle was used in each stage of the competition for each bar is highlighted. Each participant was shown both a stacked and a grouped bar plot in one of three colour schemes; the default for the language, viridis, and greyscale. For three versions of the survey, the stacked bars were shown first, and for the other three versions the first shown was the grouped bars. The final question of this part also asked respondents to compare two colour schemes, and through the 6 surveys there are comparisons of every colour scheme against every other colour scheme.

The question *"How many times would you say 'Floating Steps' were used in the Finals (Regional/City) round?"* is the first here, and is regarding the reading of a numerical value off the axis. In this question respondents were asked about 'Floating Steps', which is the bar third along from the y-axis. The question asks respondents to view the bar plot, where the bars will either be grouped of stacked, and decipher how many times this obstacle was used in the specified round of the competition. The true value for this was 11. The hypothesis for this question is that the respondents will more accurately gauge the value for the grouped bar than the stacked, which as seen below appears to be the case.

The mean for the values estimated by respondents using the stacked bars is 14.32, a fair bit larger than the true value of 11, and the mean estimated value for the grouped bars was closer to the true value, at 11.8. The IQR for the grouped bars is also smaller than for the stacked, and comprises of the range $[11, 12]$, insinuating that the estimated values tended to be fairly accurate but with some respondents perhaps slightly overestimating. The IQR for the stacked bars on the other hand covers the interval $[10, 14]$, which does contain the true value, but shows a tendency for both over and underestimation of respondents. Additionally to this, there is a large variance in the responses to this question, at 54.8 compared to the variance of 13.1 for the responses regarding the grouped bar plots. This adds to the picture that there was much less agreement between respondents, with many straying away from the mean of 14.3. It is seen however that the median for both the stacked and grouped bars is 11, showing that the higher mean of the stacked bars may be a result of an influential value at the upper end of the distribution, and that many observations do actually sit around 11. The fact that many values actually sit around 11 could be contributing to the higher variance, as variance is simply the sum of the squared distances from the mean, and so will be elevated if there are many values that sit some distance away from the mean. The higher mean could be reflected in the maximum of the stacked responses being 35, although the maximum of the grouped responses is 40, so there may be more than one influential point in the stacked responses.

Outliers can be checked for by looking at the box plots for this data.



It can in fact be seen that the box for the grouped responses is short and centered around 11. The box for the stacked responses shows many high valued outliers that could be causing the mean to be higher, although the IQR is still a fair bit larger than that of the responses for the grouped bars. The mean for this also sits above the IQR, and thus the outliers may be having a significant influence. Now the outliers will be removed, assuming, from the box plot, that outliers are any values above or equal to 25 for the stacked responses and above or equal to 20 for the grouped.

Removing the outliers as specified by the box plot, the mean of the stacked responses is now just above 11, and actually closer to the true value than the mean of the other set of responses, and the median has decreased to 10. From this one could infer that there is no difference between each type of bar plot in terms of gauging the size of the bars. However, there are 12 outliers in the stacked responses, which leads to the idea that these are not in fact all outliers and may be valid responses that just sit on the upper end of the distribution. However, it seems the cause of the high values could be respondents taking the whole height of the bar, which has an actual height of 28, rather than the section of interest. Many of the potentially influential values fall around the range $[25, 30]$, with all but 2 of the 12 potential outliers sitting in this interval, with the remaining two both being 35. Looking below at the summary statistics for only the values picked up as outliers, there is a mean of 29.83, which is higher than the true value of 28, and interestingly goes against the analysis from part 1, question 2 whereby respondents were asked to judge the height of this bar and on average underestimated. The fact that so many participants misinterpreted this plot and signify that stacked bar plots may not be the best way to present data to general public, as there may be the potential to misread the height of the whole bar as the size of the top category.

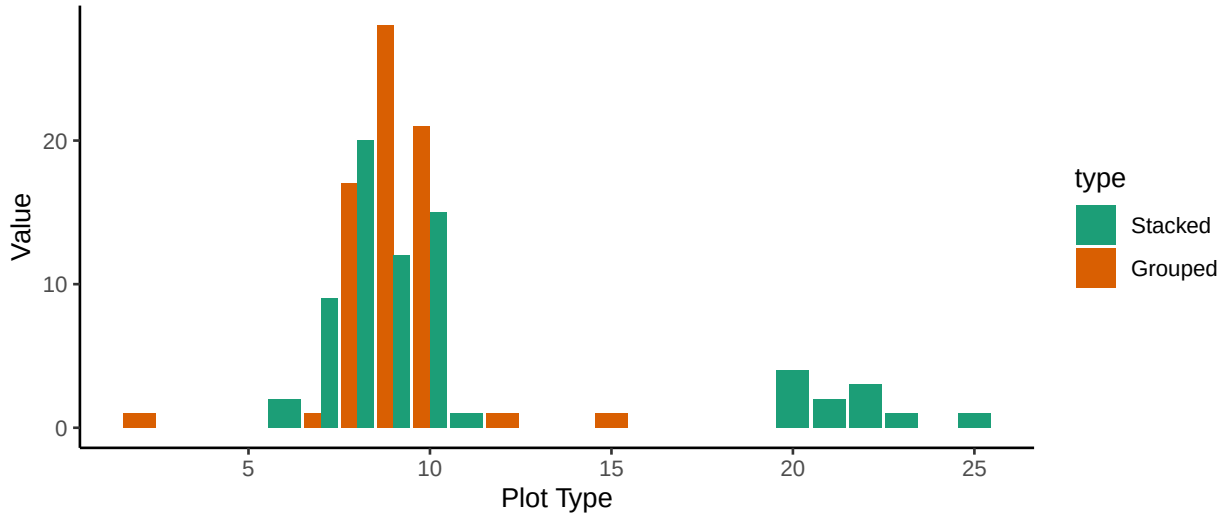
As a result of this, this set of 12 values will be discounted from the analysis, and thus come to the conclusion that, for the respondents that appear to have judged the height of the correct section,

there was little to no impact when using stacked vs grouped bar charts, and most of the difference comes from misinterpretation of the plot itself, as opposed to a poorer judgment of size.

To see if either of these values are significantly far from the true value, tests are once again run. A sign test on the stacked bar plot responses gives a high p-value of 0.5258, showing that for the stacked bar plot responses (after removing the values as priorly specified), the participant estimated values do not differ significantly from the true value. For the grouped bar plot the obtained p-value is $0.009 < 0.05$, and thus these responses are statistically significantly different from the true value. Running t-tests on the means, however, sees both sets of responses differing statistically significantly from the true value.

The next question, *'How many times would you say 'Log Grip' was used in the Finals (Regional/City) round?'*, is similar the above, but for the next bar to the right. The purpose of this question was to test the same hypothesis as the previous question, and also to lead into the following question, where respondents were asked to compare the 'Floating Steps' and 'Log Grip'. Additionally, the bar in the previous question had only two categories, of which the respondents were asked to judge the size of the category on the top of the bar in the stacked plot, whereas the bar for 'Log Grip' has 5 categories, of which the category of interest sits above 4. The true value of this was 9.

Similarly to the previous question, the mean response for the stacked bar plots are higher than that of the grouped, and the mean of the stacked also slightly overestimates the value. Once again however, a selection of respondents appeared to judge the full height of the bar rather than the category as asked.



Indeed, the distributions of values for each of the two response sets appear to be almost identical

After removing the outlying values, there tended to be a slight underestimation in the value for the stacked bar plot, however this is approximately 0.46 away from the true value, and unlikely to be

significant.

Once again the response sets are non-normally distributed and asymmetric, and so sign tests are applicable. The response set for the stacked bar plots produces a p-value of around 0.04, which shows a statistically significant difference in the responses from the true value of 9 at the 0.05 level of significance. However, this would easily become insignificant by slightly lowering the significance level to, say, 0.035. The p-value for the grouped bar responses, however, is » 0.05, as expected given that the median of the data sits at the true value.

The t-tests show that the differences in the means from the true value are statistically significant, although not considering the tests it can be seen by eye that the means are relatively close to 9.

The respondents were then asked to *'How many times would you say 'Log Grip' was used in the Finals (Regional/City) round?'*. This question asked respondents to judge whether log grip was used more, less, or an equal amount in the Finals (Regional/City) and Qualifying(Regional/City) rounds. This was to see how well differences between sizes of categories are judged when relating to the same variable, and are in the same bar. The results for this are given in the table below.

The table shows overwhelmingly that significantly more people accurately judged that the two values were the same for the grouped bars than for the stacked bars. This was the hypothesised result, and has presented to an even greater extent than previously anticipated. All but 7 of the respondents who responded to this question correctly judged from the grouped bars that the obstacle was used an equal number of times in each of the two rounds, whereas the responses for the grouped bar seemed fairly well split between the three options. It may be interesting in the multivariate analysis section to compare responses depending on whether respondents were shown the stacked or grouped bars first. Perhaps a reason for the incorrect judging with the stacked

Respondents were then asked *'Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?'*Similar to the previous question, this asks for a comparison between the size of two categories, but this time about how many times two different obstacles were used in the round Finals (Regional/City), where these two obstacles are those discussed at the start of this part of the survey.

This was a potentially poorly formulated question, as the respondents had already been asked to specify how many times each of these obstacle was used in this round and respondents mostly judged this accurately with regard to both plots, but this could have been impacted by the previous questions. However, this does follow from the results from the past questions showing that respondents mostly accurately judged the values correctly, aside from those who instead judged the height of the whole bar.

The aim of the question *'Which bar chart do you feel is easiest to read and interpret?'* was to assess

the perceived ease of interpretation of both bar plots. This is to gain an understanding in how data may best be presented in an easily understandable, easily readable manner. This is an important factor in visualisation, as a main aim in creating visuals is to provide an aid for the viewer to simply and quickly see the message. The opposite may be beneficial in certain applications however; based on the misreadings in the question regarding judging the number of times 'Log Grip' was used in the specific round, viewers of the visualisations could be easily mislead by incorrectly interpreting the plot. The people being shown the plot in, for example, an advert, may only take a fleeting look and not go beyond to analyse the plot to see accurate differences between values, and thus it is important to produce a plot that gives the easiest interpretation.

Var1	Freq
Grouped	59
Stacked	11

The large majority of participants found the grouped bar chart easier to read and interpret, as predicted.

The questions *'Which bar chart do you feel is easiest to read and interpret?'* and the one following *'Do you feel that one of the colour schemes makes it easier to read and interpret? If so, please select which one.'* are asked with the purpose of assessing the colour scheme that gives the greatest aesthetic pleasure, or effectively which colour palette the respondents feel is subjectively the 'prettiest' or 'nicest'. It is important to note here that aesthetics and readability do not always go hand-in-hand; a plot that is made to look very aesthetically pleasing may sacrifice readability, and vice versa. For each of the two languages, six pairings of three different colour palettes were created, whereby the first colour was the one displayed for the main questions, and the second used only for the comparison questions. As previously discussed, the three colour schemes considered are viridis, greyscale, and each language's default plotting colour palette. The colour palette pairings are outlined below, where each set of two colours is assigned a 'Pairing ID' from A to F.

Pairing ID	Main Palette	Secondary Pallette
A	Viridis	Default
B	Default	Viridis
C	Default	Greyscale
D	Greyscale	Default
E	Viridis	Greyscale
F	Greyscale	Viridis

This table shows that when it came to the default/viridis pairings, displayed in the first two rows, the respondents tended to have no preference overall, although this may differ between languages, which will be explored later on. Comparing this to the bottom two rows, in which viridis is put against

	A	B
Set A	7	6
Set B	6	6
Set C	9	1
Set D	3	9
Set E	11	0
Set F	1	11

greyscale, only 1 respondent out of the 23, a proportion of 0.04, found the grey more aesthetically pleasing, as hypothesised. When considering greyscale/default, there was still a majority preferring the non-greyscale palette, but a higher proportion preferred this as compared to the viridis/greyscale, with 4 out of the 22, or a proportion of 0.18, preferring the grey.

Complementing the aesthetic preferences, the second question assesses the colour preference with regard to readability and ease of interpretation. As mentioned before, this will be used to test both the colour palette preference itself alongside whether this preference matches up with aesthetic preference.

Var1	Freq
A	42
B	20
None	8

Interestingly here, the top two rows appear to give opposing results; the respondents who were presented with viridis for the main questions and the default as a secondary palette stated that they found either viridis easier to interpret or had no preference, whereas those presented with the default first and viridis second tended to find the default easier. Once again looking at the comparisons with the greyscale, there were some respondents that found this easier to read, but the majority chose the alternative, whether this is viridis or the default.

Sales - Part 1

Now consider the sales part of the survey. In this section data was taken from a the `BJsales` data set in R, which is a time series data set containing 150 observations. This data set constitutes a single vector of values with no specified timings, and the visualisation data was formed by taking subsets of size 12 this and setting a month between each point to give a year of fictional sales data.

How much would you say sales of each company increased between January and December? [Company A]

This question was included for the purpose of testing whether, again, axis scaling impacts the perceived differences between values, but this time with time series line plots as opposed to bar plots. Respondents were asked to assess how much the sales of company A increased over the course of the year, or in other words to look at and compare each end of the line.

The plot for which the respondents, on average, found the difference to be smallest was the zeroed, followed by the truncated, and then the separated, with means of 1.371, 2.414 and 3.043 respectively. These differences are found to be statistically significant, as outlined in table[?].

Table 3: Table of p-values for this question

Alternatve Hypothesis	P-value
Truncated > Zeroed	8.870681966755e-14
Truncated < Separated	0.00654175643803223
Separated > Zeroed	3.48079934270661e-13

The differences between languages and plot ordering were shown to be inconsequential (see table [?])

How much would you say sales of each company increased between January and December? [Company B]

The zeroed was once again perceived to have the smallest difference ($\bar{x} = 1.371$), but this time with the separated in the middle ($\bar{x} = 4.1304$) and truncated with the largest difference ($\bar{x} = 4.1304$). See table [?] for p-values. The p-values show sufficient evidence that the truncated responses were on average greater than the zeroed, as were the responses for the separated plots. However, the difference between the ratings for the truncated and separated plot responses was inconsequential, along with the language comparisons and plot order.

Table 4: Table of p-values for this question

Hypothesis	P-value
Truncated > Zeroed	8.95254768631571e-23
Truncated not equal to Separated	0.2162
Separated not equal to Zeroed	12.46327564235365e-23

How large would you say the drop in sales between April and July of Company A is?

The means for this question appear very significantly different by eye, once again with the zeroed plot eliciting the lowest average rating ($\bar{x} = 1.371429$), followed by the truncated ($\bar{x} = 2.814286$) and then the separated ($\bar{x} = 4.028571$). The p-values confirm the significance of the differences between all three variables.

Table 5: Table of p-values for this question

Hypothesis	P-value
Truncated not equal to Zeroed	1.03832498155043e-11
Truncated not equal to Separated	0.00012743463393642
Separated not equal to Zeroed	1.1261341031207e-16

Sales - Part 2

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?

The final question of the survey compares just two plots, for which the difference in the ratings is shown to be significant, with the mean for the truncated plot ratings at $\bar{x} = 4.271$ and for the zeroed $\bar{x} = 2.7$ and a one-sided p-value of $p = 4.44089209850063e - 15$ showing the difference in the truncated was on average rated as larger than for the zeroed.