Worcester Polytechnic Institute
Digital WPI

Major Qualifying Projects (All Years) | Major Qualifying Projects

April 2018

Human Perception of Outliers in Correlated Scatterplots

Tyler K. Jaskoviak
Worcester Polytechnic Institute

Human Perception of Outliers in Correlated Scatterplots

A Major Qualifying Project
Submitted to the Faculty of
Worcester Polytechnic Institute
in partial fulfillment of the requirements for the
Degree in Bachelor of Science
in
Computer Science and Psychological Science
By

Tyler Jaskoviak

Date: 4/25/18

Project Advisors:

Professor Lane Harrison, Advisor

Professor Jeanine Skorinko, Advisor

Abstract

Despite ample research analyzing how people recognize differences in data, one aspect that has largely gone unmeasured is how outliers affects these comparisons. This paper aims to provide a better understanding how people recognize differences in data by having participants decide which correlation is stronger (forced choice) when comparing scatterplots at different correlations with outliers. With 67 participants, we calculated a just noticeable difference (JND) at different correlation values. The results indicate that at all levels of correlation (e.g., .4 or .8) tested, people were less able to detect the stronger correlation for the scatterplots with outliers compared to scatterplots without outliers.

Intro

In our everyday lives, we are likely to see forms of data and be expected to draw conclusions from these different data formats. Often, we hear news stories making a claim and showing a chart that supports that claim. But, it is up to the viewer to ultimately make conclusions about the data.

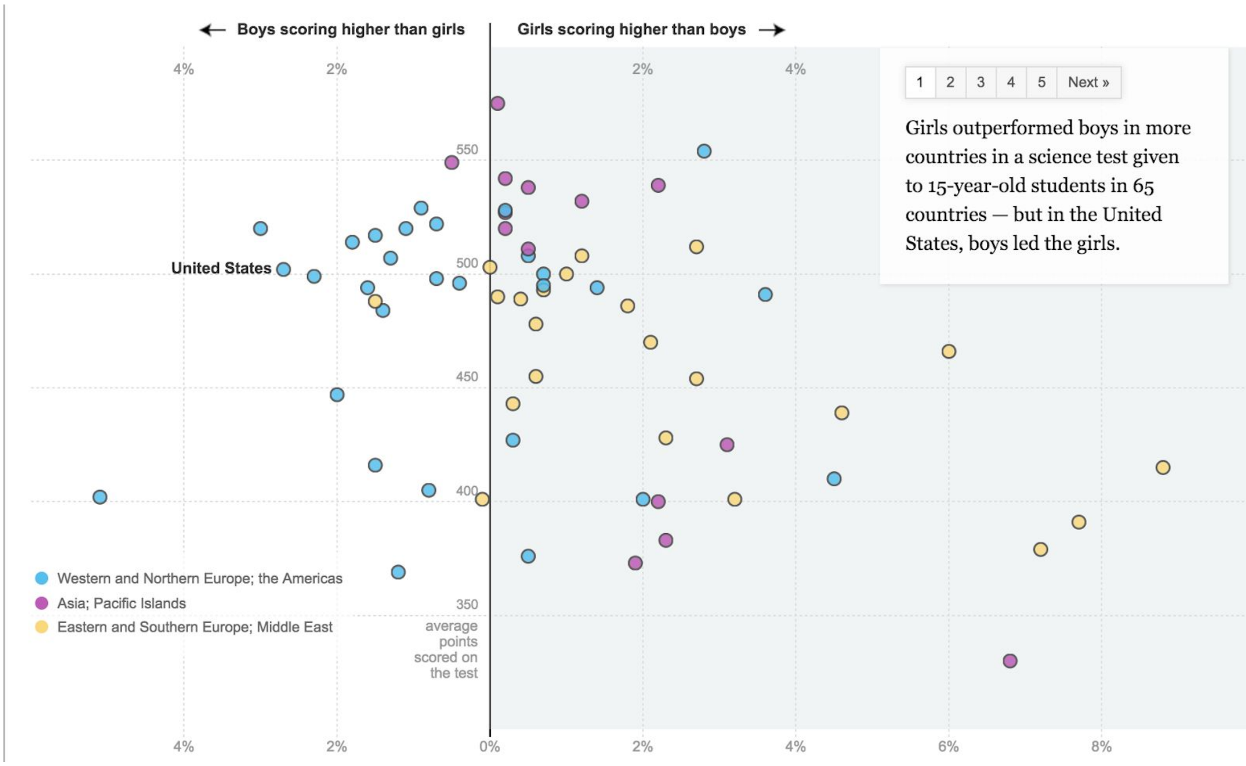


Figure 1: A scatterplot published by the New York Times in 2012. Each data point represents a country. The x-axis is which gender performed better on a sponsored test, left indicates that boys did better, right indicates that girls did better. The y-axis indicates the average score that was scored on the test.

For example, the above figure was published by the *New York Times* in 2012 and reports on how countries performed on a science test provided by the experimenters. The assertion of the article (and scatterplot) is that in certain regions girls performed better than boys and this can have cultural reasonings. This is seen by all the data points on the right side of the origin. Additionally, there is a general trend moving to the bottom right which could indicate that there is a positive relationship with the percentage of girls who outperform boys and the average score on the test. One factor that may influence how people interpret this graph is that there are a number of datapoints that do not follow the general trend of the rest of the data. For example, the blue datapoint towards the bottom left of the data is for Columbia. This data point goes against the trend of the data. The question then is how do researchers, reporters, and others who view this graph treat these data points that do not follow the general trend? Do viewers exclude them from their analysis? Do viewers take these points into consideration? Do these outlying data points influence how strong the relationship appears to be? The current research aims to address these types of questions by determining the effect that outliers can have on interpreting the data.

Detecting Differences in Graphs

Data and visualization research has recently focused on looking at how people react and understand general trends. In an experiment by Rensink and Baldridge(2010) as well as by Harrison, Yang, Franconeri, & Chang, R (2014), the experimenters had participants compare two scatterplots at different correlations in a forced choice response. The purpose of this was to determine the minimum difference between correlations that participants could distinguish as being different using just noticeable difference (JND). JND refers to the minimum difference between a stimulus that is still noticeable by human perception (Stern, 2010). The conclusion from both studies indicates that as the correlation of a base scatterplot increases then the JND decreases. JND can then be directly mapped by Weber's Law:

$$\frac{(JND)dS}{S} = constant$$

which states that given a stimulus, there is a linear relationship between the stimulus intensity and the JND for a given difference in the stimulus (Rensink & Baldridge 2010). This research gives a basis of what we understand about visualizations and different correlation values. We aimed to expand this research by looking at introducing outliers to the datasets being compared.

Outliers

Outliers are classified as data points that do not follow the general trend of the data (Hoaglin, 2003). Beyond this common definition there are several mathematical definitions of an outlier. Most prominently used is Tukey's Fences which defines an outlier as any value 1.5 times the Interquartile range (IQR) above the third quartile or 1.5 times the IQR below the first quartile as seen in Figure 2 (Hoaglin, 2003). These values are most commonly used in boxplots.

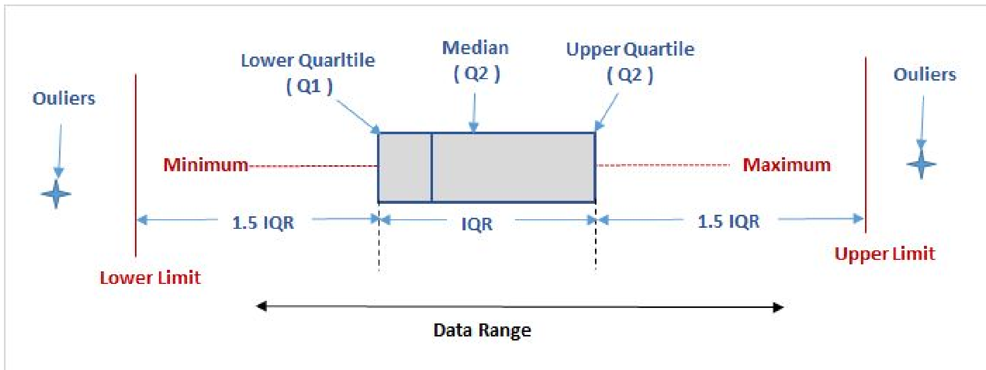


Figure 2: An indication of how outliers are defined by Tukey's fences. The IQR is used to create an upper and lower limit for a given boxplot. Points that lie beyond this limit are considered outliers.

Another, algorithm that has shown popularity is Chauvenet's criterion (Cattel, 1903) which defines outliers as points that do not fall within an allotted deviation:

$$D_{max} \geq \frac{|x - \mu|}{\sigma}$$

where *Dmax* is the maximum deviation that points may fall upon, *x* is the value of the point, μ is the sample mean, and σ is the sample standard deviation. Any result that is larger than *Dmax* is considered an outlier.

Outlier detection is also a common test for computer algorithms data based on different criteria as demonstrated by Kriegel et al. (2010). One of the most common computer definitions is k-nearest neighbor which gives an indication where each point would be able to map the *x* closest data points to itself. Outliers can be detected with this method by finding that the distance between its neighbors is some limit larger than all other points as seen in Figure 3.

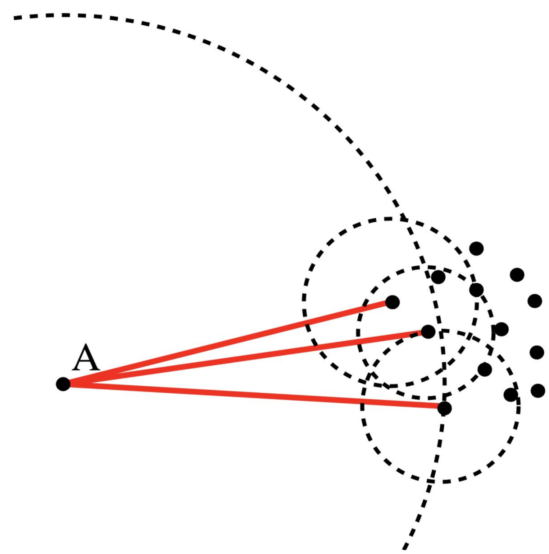


Figure 3: An example of outliers in k-nearest neighbor. Point 'A' is shown to be an outlier because the distance to its three nearest neighbors are significantly larger than that of the other points.

This can be modeled mathematically into the following equation for using distance to calculate outliers:

$$ROF(p) = \sum_{R\min \leq r \leq R\max} \frac{clusterSize_{r-1}(p) - 1}{clusterSize_{r}(p)}$$

where *Rmin* and *Rmax* are the minimum resolution and maximum resolution respectively. This translates such that *Rmax* is the resolution maximum which translates to the minimum distance threshold that something has to be over to be considered an outlier and *Rmin* is the resolution minimum which translates to the maximum distance threshold where everything is considered part of the cluster. This summation yields a number where every point has been considered both an outlier and part of the cluster and the score reflects which group the point belongs to.

Current Research

The purpose of the current research is to expand what we know about data visualization human perception by examining how outliers influence data visualization. To do this, the same experiment performed by Rensink and Baldridge (2010) and Harrison and colleagues (2014) was conducted except the graphs included outliers. The reason for this is to determine if performance decreases with more complex data. The current work will increase our understanding of how outliers can affect our perceptions of data and visual representations of data (i.e., graphs). Furthermore, the current work may provide insights into different practices that should be used when displaying data with outliers, especially to a general public audience.

Method

Participants

A total of 67 individuals participated in the study. Sixty individuals participated through Amazon's Mechanical Turk, and 7 participants were undergraduates from a private institution in the northeast portion of the United States. Participants from Mechanical Turk were paid $3.60 for their time in accordance with estimated time and federal minimum wage, and undergraduate participants earned course credit or volunteered their time. All tablets and mobile devices were blocked from the study in order to avoid confounding variables. Of the participants, 36 were male and 31 were female and the median age range was 25-31 years.

Terminology

To avoid confusion, we wanted to define some of the terms that we use in this section. A *trial* is when a participant makes a single comparison between two scatterplots, an example of this is displayed in Figure 4. A *run* is a group of trials that all share the same independent variables. A *round* is a group of runs that participants completed that include all the different independent variables. For this experiment, there were two rounds: the practice round and the test round.

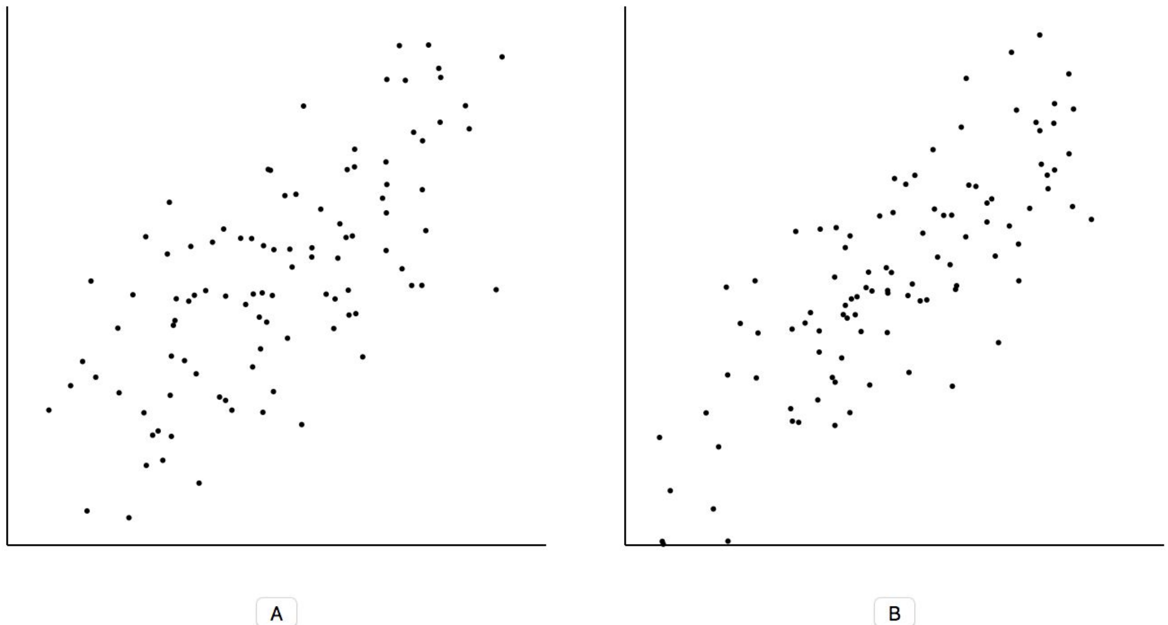


Figure 4: Example of one trial. A participant would compare two scatterplots like above and select which ones is more highly correlated. In this example, the left correlation has an r = 0.8 and the right correlation has an r = 0.9 meaning that the right scatterplot is more highly correlated and the correct answer.

Materials

Stimuli presentation. The stimuli were modeled from past work conducted by Rensink and Baldridge (2010) and Harrison and colleagues (2014). Each scatterplot was 300 x 300 pixels, contained 100 data points distributed normally along the 45 degree line with a pixel size of two. Additionally, the bottom and left axes were displayed.

Data generation. Data was generated using the Harrison and colleagues (2014) method that was modified from Rensink and Baldridge's (2010) equations. Using r_z to denote the correlation coefficient of the dataset after generation, each point (x_i, y_i) is transformed using:

$$y_i' = \frac{\lambda x_i + (1-\lambda) y_i}{\sqrt{\lambda^2 + (1-\lambda)^2}}$$

where λ is defined as:

$$\lambda = \frac{(r_z - 1)(r^2 + r_z) + \sqrt{r^2(r_z^2 - 1)(r^2 - 1)}}{(r_z - 1)(2r^2 + r_z - 1)}$$

This is using the modified equation for λ that was developed by Harrison and colleagues (2014) because it converges more quickly and eliminates the error of ± 0.005 from Rensink and Baldridge's method (see Harrison, et al., 2014).

Outlier generation. Since previous work has not examined outliers, we developed a method to create these outliers. To create outliers, we selected 5 data points at random in a given plot and randomly distributed them into an ellipse of size *r1*= 25 and *r2* = 13 pixels. We then placed the points along the minor axis of the rest of the data and moved them 3 standard deviations from the center either above or below the main plot as seen in Figure 5. Next, the outliers were then rotated to a 45 degree angle to match the style of the rest of the data. Finally, the rest of the data, not including the outliers, were readjusted by iterating λ from the Harrison and colleagues (2014) equation by 0.0001 until the final correlation was larger than $r_z - 0.0000001$ or λ was greater than 0.99999. This final change in stimuli resulted in a dataset where the correlation of the entire set, with the outliers, was equal to the target correlation. The effect, after all these steps, is the same data formed by Harrison and colleagues (2014) and Rensink and Baldridge (2010) but with 5 outliers in a cluster, 3 standard deviations along the minor axis of the rest of the data.

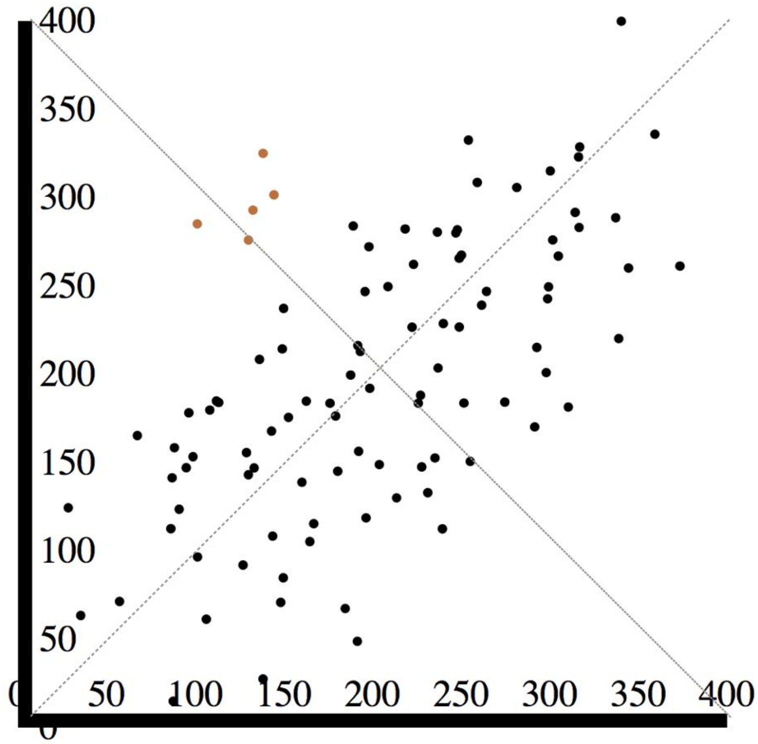


Figure 5: Scatterplot with modified points. In dotted lines is the major axis, in the direction of the data, and the minor axis, perpendicular to that data.

Base correlation manipulation. One independent variable was the the base correlation (r = 0.4, 0.6, and 0.8). For each comparison that a participant made, one of the two scatterplots remained at the same correlation while the other correlation became closer to the base with a correct answer and further away with an incorrect answer. For example, when comparing a base correlation of 0.8 to a variable correlation of 0.9, the 0.8 will remain fix for the duration of the run and the variable correlation will change based on performance.

Approach direction manipulation. For each trial, participants viewed a scatterplot that had the base correlation and a second graph whose correlation varied based on the participant's correct (or incorrect) responses. The second independent variable in this study was the direction of the scatterplot (above or below) for the non-base or variable correlation. Above indicated that the variable correlation was higher than the base correlation (e.g., base correlation = 0.8 and non-base/variable correlation = 0.9) and below indicated the opposite (e.g., base correlation = 0.8 and non-base/variable correlation = 0.7).

Outlier placement manipulation. Another independent variable manipulated which of the two charts included an outlier (higher or lower). Higher indicated that the scatterplot with the higher correlation included an outlier while lower meant that the scatterplot with the lower of the two correlations included an outlier. In all cases, only one of the two scatterplots had a cluster of outliers.

Procedure

After reading the informed consent and providing a unique user ID to prevent multiple submissions, participants viewed a screen that explained what a correlation was and they also viewed a set of example correlations ranging from 1 to 0.1. Next, participants completed a round of practice problems. In these practice problems, participants saw two scatterplots that did not include outliers. Participants then indicated which of the two plots had a stronger correlation value. Participants learned if their answer was correct. If participants answered correctly, then the program made the correlations closer together for the next trial (e.g., from $r = 0.8$ and $r = 0.9$ to $r = 0.8$ and $r = 0.88$). If participants answered incorrectly, then the correlations were further apart for the next trial (e.g., from $r = 0.8$ and $r = 0.85$ to $r = 0.8$ and $r = 0.87$). This scaffolding procedure based on the correctness of the answers was used in Harrison and colleagues(2014). Participants completed two practice runs of 15 trials each with a break in between.

After completing the practice round, participants completed the test round. In the test round, participants completed the same task; however, this time one of the two scatterplots included outliers. Participants completed 12 runs, one for each condition set by the three variables: base correlation (0.4, 0.6, 0.8), approach direction (above, below), and outlier placement (higher, lower). Participants compared each round for up to 50 trials depending on how quickly the JND was determined. Each participant was given a short break after each run. After completing the study, participants completed demographic information including age and sex. Participants also indicated their familiarity with correlations and scatterplots. Participants were then debriefed.

Results

We found several significant results upon analysis of our data. We ran a one way ANOVAs (Analysis of Variance) for each independent variable on the mean JND of that group.

Base correlation. We found that there was a significant difference between the JNDs all of our base correlations levels with an $F(2, 801) = 78.209$, $p < 0.001$ (Table 1) after running an Tukey HSD Post Hoc analysis, giving an indication that this model follows Weber's Law. This is the exact same result that the previous studies also found.

Base Correlation	*M (JND)*	*SD*
0.8	0.05980	0.07054
0.6	0.10585	0.09031
0.4	0.16659	0.12790

Table 2: The mean and standard deviation of each base correlation level relative to JND. The higher the correlation the lower the JND which works as Weber's Law predicts. It was found the difference between each result was significant at the $p < 0.001$ level.

Approach direction. Additionally when comparing the approach direction (above, below) compared to JND, we also found that there was a significant difference at $F(1,802) = 19.178$ $p < 0.001$ (Table 2) however, we found that this effect works counter to Weber's Law. We would expect that the above approach would yield a lower JND as the the two correlations would be higher and therefore easier to recognize the difference between, but instead the results indicate that it is easier to recognize the difference between the correlations when the below approach was used.

Approach Direction	*M (JND)*	*SD*
Above	0.12728	0.12304
Below	0.09421	0.08823

Table 2: The mean and standard deviation difference for the above and below approach. It was found significant between the two conditions at $p < 0.001$.

Outlier placement. We found that between conditions, there was not a general effect given by the placement (higher or lower) of outliers ($p = 0.094$). Since the p-value was marginal in nature, we conducted an exploratory analysis to examine if results were significant when running a two-way ANOVA at different values. We found that there was a significant result only at the 0.4 base correlation condition as seen in Table 4.

Base Correlation	Outlier Placement	*M (JND)*	*SD*	*p*	r^2
0.4	Higher	0.14716	0.012147	0.013	0.023
	Lower	0.18601	0.13162		

Table 4: The outlier placement had a significant effect at the 0.4 base correlation level with a $p = 0.013$, this gives evidence towards increased performance when the higher of the two plots has an outlier.

Participant sex. In relation to the sex of the participant, we found that female participants performed significantly better than male participants ($F(2,108) = 11.480$, $p < 0.001$). This difference can be observed in Table 5 below.

Sex	*M (JND)*	*SD*
Male	0.12648	0.13146
Female	0.09160	0.06901

Table 5: Female participants averaged a lower JND indicating higher performance over male participants ($p < 0.001$).

Previous visualization experience: Participants previous data visualization experience significantly influenced their performance. Participants who either had very little previous visualization experience (i.e., "1") or extensive previous visualization experience (i.e., "5") performed more poorly on the task than those with less extreme previous visualization experiences, ($F(4, 799) = 7.460$, $p < 0.001$). This can be seen in table 6. We additionally ran a Tukey HSD Post Hoc analysis and found that $p = 0.001$, $p = 0.015$, and $p = 0.012$ when comparing an experience rating of 1 to 2, 3, and 4 respectively. When comparing an experience rating of 5 we found that $p = 0.001$, $p = 0.008$, and $p = 0.004$ relative to 2, 3, and 4 respectively.

Visualization Experience	*M (JND)*	*SD*
1 (very little)	0.13265	0.13803
2	0.08980	0.07766
3	0.10199	0.09722
4	0.09061	0.06523
5 (very much)	0.14940	0.12737

Table 6: It was found there a significant difference between an experience of 1 relative to 2, 3, and 4, in addition to a difference between 5 relative to 2, 3, and 4. This indicates that having a lot of experience as well as no experience both decrease performance.

Other demographic variables. With respect to all other demographic variables there were no other significant results, this includes for age (p = 0.056), handedness (p = 0.990), and experience with correlations (0.910). We did find a significant result with monitor size where smaller and larger monitors indicated better performance than middle size ones (17''-19'') ($F(7,796) = 16.442$, $p < 0.001$), however, the largest n for the conditions is 14 therefore a larger sample would need to be collected for more definite conclusions.

Compared to past experiments. In order to analyze the data relative to if the participants did not see outliers we compared our data to that of one of the previous studies (Harrison et al.,2014). We found that on each level of correlation, there was a statistically significant difference between our data and that of the previous study as seen in Table 7. This indicates that when outliers are included in the scatterplots, they are more difficult to compare as the JNDs are higher in those cases.

Base Correlation	*M*, our data (*JND*)	*SD*, our data	*M*, past data (*JND*)	*SD*, past data	*F(1,324)*	*p*	r^2
0.8	0.05980	0.07055	0.03923	0.01368	9.192	0.028*	0.015
0.6	0.10585	0.09031	0.08180	0.06211	3.953	0.047*	0.012
0.4	0.16659	0.12790	0.11444	0.05971	4.877	0.003**	0.028

Table 7: Comparison of our data to that of a previous study (Harrison et al., 2014). A lower mean indicates greater ability to detect the difference between two scatterplots. * is significant at the 0.05 level, ** is significant at the 0.01 level

General Discussion

The current research set out to investigate how individuals perceive data (namely scatterplots and correlations) that contain outliers. Overall, the results from this study indicate that individuals have a harder time interpreting data that has outliers in it as the just noticeable difference (JND) is greater than when data has no outliers. Our results also provide further evidence that this type of experiment follows Weber's Law relative to base correlation. However, we found that the approach was counter to what we would expect based on past work (Rensink et al. 2010; Harrison et al. 2014). More specifically, past work found that when both scatterplots had a higher correlation (i.e., above approach) it was easier to identify the JND. However, in the current work, participants had a better JND when the lower correlation was variable (i.e., below approach). This result is could be an indication of how outliers can make the visualization process more complicated. . Future research should continue work in this area to see if this pattern is replicated. .

Contrary to our predictions, the outlier position did not significantly influence JND. One potential reason for this lack of significance may be that the difference in the correlations between the higher and lower conditions was not great enough to yield more significant results. Future research should continue to explore outlier position and the effects it may or may not have on JND and data visualization.

Our exploratory analysis on participant sex indicated that female participants performed better on the task relative to male participants. This could be connected to past research in which women perform better on JND and other sensory specific tasks (Williams, 2015). Past research has also connected this to age, however, we do not have a large enough sample size at the older age ranges to make a comparable conclusion. Again, future research should continue to examine the effects that participant sex and age have on data visualization and JND.

One perplexing finding we had was that with participants previous visualization experience. Our results indicate that those with very little but also those with very extensive previous visualization experience performed worse on the task than those with less extreme previous experience. This finding may be an indication of an optimal level where having some experience improves performance but having too much, in either direction,decreases it. Having decreased performance with less experience makes sense as these individuals are very new at the task at hand. However, we were not anticipating that those with extensive experience would perform at the same level as those with very little previous experience. It is important to note that he decreased performance with high experience could be because of the small sample size in that condition. This could be something to look further into in future experiments.

The result we found to be most telling of our experiment is the comparison between our data and past data. This was done because both experiments were built upon the same code base and there would be an increase risk of fatigue if our experiment increased in length. The results of this comparison give evidence towards decreased performance when one of the scatterplots under comparison has outliers. This can be extended to indicate that when the dataset has more complexities there is decreased performance by those viewing it. Further research could look at different ways to add different kinds of complexities to graphs to see if there is any difference, this could include more clusters of outliers, different size clusters (both in number of points and spacing of points), as well as combinations of these.

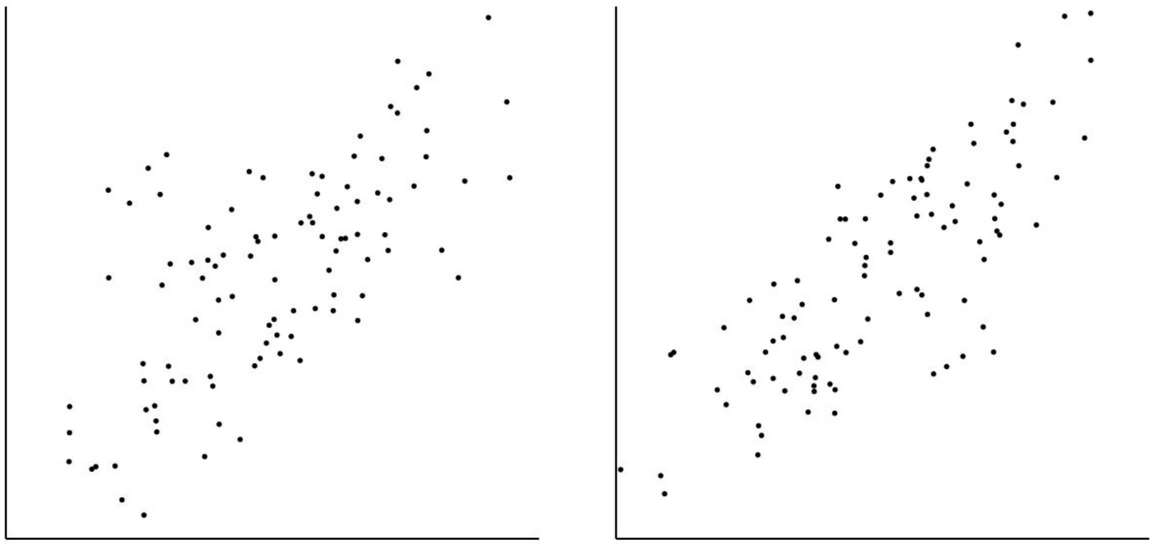
Because of the results found in this experiment, we believe that further research into human perception of outliers is needed. While the current work indicates that outliers make detecting correlations more difficult than when outliers are not present in the data, future research into outliers would provide a better idea of what different outlier conditions make this task more difficult or if there is a point where the task becomes easier. Additionally, further experimentation could yield evidence as to why these differences occur.

Limitations and Future Research

This research was limited largely by length of time it takes to complete given the scaffolding method employed and the potential for fatigue on behalf of the participant. Therefore, we only collected data comparing one scatterplot with outliers to a scatterplot without outliers. If we collected data where both scatterplots did not have outliers, it would have increased the power of our data to allow us to detect any differences when completing both tasks at the same time. However, we opted to simplify the design and only look at the comparisons of scatterplots with outliers to those without outliers to reduce chance of fatigue in the participants.

Additionally, while we randomized which side the outliers would appear on (the upper-left or lower-right relative to the data), we did not collect data on the ratio for how many outlier clusters appeared in each position. Thus, a participant could have received most scatterplots that had outliers in one position. We are unable to determine if this could have influenced the data in anyway. Future research should look into this. .

In addition to the future directions already discussed, there are many other potential directions that this research could go in the future. One direction is to have both charts include outliers instead of the one as was done in the current study. This might cause even greater decreases in performance as the data would increase in complexity. However, it is possible that by viewing two scatterplots with outliers might help performance since there would be complexity in both graphs. Additionally, this study could be run with negative correlations instead of just positive ones.



A B

Figure 6: Comparison of a base correlation of 0.8 with outliers above the data relative to a variable correlation of 0.9 with outliers below the data.

There could also be a potential result when the outliers are above or below the data (see Figure 6 for an example). Past research has suggested that people have a preference for either the left or right side of a given image and this has potentially cultural factors (Shaki, 2012).

Finally, similar to the Harrison and colleagues(2014) experiment, we could run this experiment again but with different types of visualizations other than scatterplots. This could include parallel coordinates, stacked area, and even donut charts.

Conclusion

The purpose of this study was to examine how outliers influenced perceptions of data. In this particular study, we examined how outliers affect participants visual ability to determine the strength of a correlation. Based on the results from this study, we conclude that when outliers are present then there is a decrease in performance because the outliers increase the complexity of the visual stimuli (i.e., scatterplots in this study). Thus, the current study offers preliminary evidence of the detrimental effects that outliers can have on visualization of data. Future research should continue to explore how outliers influence visual perceptions of data to increase our knowledge of human perception and data visualization.

References

Bobko, P. & Karren, R. (1979). The perception of pearson product moment correlations from bivariate scatterplots. *Personnel Psychology, 32*(2), 313-325. doi:10.1111/j.1744-6570.1979.tb02137.x

Cattell, J. M. (1903). Statistics of American psychologists. T*he American Journal of Psychology*, *14*(3/4), 310. doi:10.2307/1412321

Fairfield, H. & McLean, A. (2012). Girls lead in science exam, but not in the United States. *New York Times*. Retrieved April 25, 2018, from https://archive.nytimes.com/www.nytimes.com/interactive/2013/02/04/science/girls-lead-in-science-exam-but-not-in-the-united-states.html?ref=science

Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1943-1952. doi:10.1109/tvcg.2014.2346979

Hoaglin, D. (2003). John W. Tukey and data analysis. *Statistical Science, 18*(3), 311-318. Retrieved from http://www.jstor.org/stable/3182748

Kriegel, H., Kröger, P., & Zimek, A. (2010). Outlier detection techniques. presentation, Ludwig-Maximilians-Universität München Universität München Munich, Germany.

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

Shaki, S., Fischer, M. H., & Göbel, S. M. (2012). Direction counts: a comparative study of spatially directional counting biases in cultures with different reading directions. *Journal of Experimental Child Psychology*, *112*(2), 275-281. doi:10.1016/j.jecp.2011.12.005

Stern, M. K., & Johnson, J. H. (2010, January 30). Just noticeable difference. Retrieved April 25, 2018, from https://onlinelibrary.wiley.com/doi/10.1002/9780470479216.corpsy0481

Williams, L. J., Butler, J. S., Molloy, A., Mcgovern, E., Beiser, I., Kimmich, O., ...Reilly, R. B. (2015). Young women do it better: sexual dimorphism in temporal discrimination. *Frontiers in Neurology*, volume 6, pages 160. doi:10.3389/fneur.2015.00160