Evaluating Visualization Styles for Likert Scale Data

Laura South David Saffo Amy Worth

Northeastern University

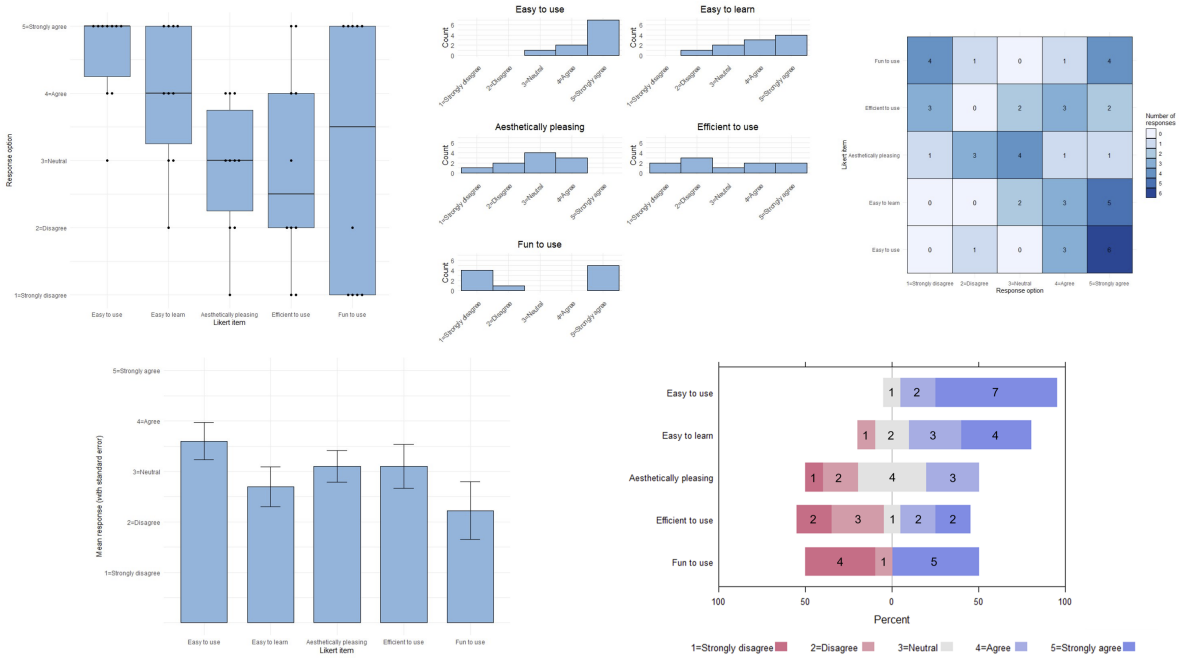

Figure 1: Example visualizations of Likert response data. From left to right: box plot, histogram, heat map, bar chart, and diverging stacked bar chart. These five styles are commonly used to visualize Likert style data in both academic papers and other forms of media. We designed each chart to be the best possible version for Likert style data. This was done be included explicit encodings or labels for the underlying distribution of the data.

Abstract

Likert scales are often used to measure subjective attributes such as user satisfaction or aesthetic value in research studies, but no empirically validated standards exist for methods of visualizing this type of data. When presenting a Likert scale response visualization authors are trying to convey a certain trend in the data. We design this study to help understand which visualization technique allows researchers to present Likert scale data more effectively. Results show visualization styles have a significant effect ($p = 0.0003$) on participant accuracy in questions related to the distributions and statistical values of the data. Out of 4 visualization styles (i.e., box plot, heat map, histogram, and diverging stacked bar chart), the heatmap stimuli was the clear outlier with higher error across all distributions. The supplemental material for this project can be seen on our OSF repository: `https://osf.io/2v6jq/?view_only=87d2895f19d5415cb5086ad81bf420ff`.

1 Introduction

A Likert scale consists of one or more statements about which participants are asked to rate their agreement by selecting one of several options, often ranging from "strongly disagree" to "strongly agree" [7]. Likert scales are often used in research to measure subjective qualities that are otherwise difficult to quantify, such as enjoyment, satisfaction, or aesthetic value.

Although Likert scales seem intuitive and simple, there are many pitfalls that researchers can encounter when constructing, reporting, and analyzing Likert scale data. For example, the number of response options, numerical representation, and phrasing can all influence how a Likert scale is perceived and responded to by participants. Once the Likert scale data is collected, researchers must choose methods to report and visualize Likert scale data, each of which can affect how the data are perceived by the reader. Similarly, when analyzing Likert scale data researchers must choose between parametric or nonparametric analysis methods based on the characteristics of their data. Empirical studies have been used to produce recommendations about best practices for constructing Likert scales [2, 4, 9] and for analyzing Likert scale data [3, 6, 8], but few recommendations exist for how to best represent Likert scale data visually.

Visual representations of Likert scale data are an important aspect of accurately and transparently reporting study results. Visualizations of this data can show clear trends from the responses, support the authors claims, and help readers understand the study in question. However, there are many visualizations that are used to visualize Likert scale data and practice, and many design decisions to be made for each one. Following conventional wisdom and best practices, a Likert scale visualization should be able to encode the trend and the

distribution of the data.

We hypothesize that visualization style will influence the comprehension of the trend and underlying distribution of Likert style responses. In this paper we contribute the results of a quantitative study comparing five visualization styles commonly used for Likert scale results. We also contribute several design recommendations and considerations derived from the results of our study.

2 Related work

Very little work has examined the best way to visually present Likert scale data. Without a full evaluation study comparing popular techniques, it is impossible to definitively state the best method for visualizing Likert scale data. Heiberger & Robbins survey a range of visualization options and recommend using a diverging stacked bar chart to represent Likert scale data [5], but no perceptual studies have been done to support their recommendation.

While studies on the best way to visually represent Likert scale data are lacking, there are many studies providing general recommendations for chart design. Cleveland & McGill conducted a study on graphical perception [1], identifying a set of elementary perceptual tasks carried out when people extract quantitative information from graphs, and ordering of those tasks based on how accurately people performed them. Studies such as this one have been used by visualization researchers to construct best practice guidelines how how certain types of data should be visually represented. These guidelines are largely what inferred our design of the five visualization styles we aim to examine in this study (figure 1).

3 Study design

3.1 Participants and Apparatus

Forty-eight participants were recruited from an undergraduate data visualization course at the Khoury College of Computer Science at Northeastern University. We used GPower to conduct a power analysis for a 5x5 two-factor ANOVA at =0.05 and found that we needed at least 35 participants to reach 80% power with a small effect size (0.25). Participants were required to complete the study as part of the curriculum for the course, thus we were able to recruit past the amount needed from the power analysis. Participants were not compensated for their time. The participants are classified as knowledgeable in the area of data visualization research.

We created 5 different visualizations 1 with styles commonly used to visualize Likert response data: bar chart, box plot, heat map, histogram, and diverging stacked bar chart. Each visualization has been constructed to follow best practices and to encode the data to the best of its ability. For each visualization style, we created 5 different underlying distributions: extreme positive, moderate positive, neutral, uniform, and split (i.e., bimodal at extremes). The encoded data on each visualization includes 10 generated responses on a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree) from each of 5 Likert items (fun to use, efficient to use, aesthetically pleasing, easy to learn, and easy to use).

We created a survey using Qualtrics which included instructions and a series of questions for each visualization style. Participants were randomly assigned through the Qualtrics survey to view the questions from one visualization style (between-subject factor), but all underlying distributions (within-subject factor) for that visualization style.

3.2 Procedure

The study took place remotely through Qualtrics, and synchronously during the course lecture. Participants were shown a chart visualizing responses to five Likert items and were asked to fill in the blanks of a pre-written "Results" section explaining the contents of the chart. Participants had to type an answer for numeric responses, but selected from several multiple choice response options for non-numeric responses. If the visualization did not provide the precise information to fill in a response, participants were directed to estimate the value to the best of their abilities. An example response section is shown below:

Ten participants were asked to complete a self-reported questionnaire using a **[numeric response]** -point Likert scale. The questionnaire measured participants' level of agreement towards **[numeric response]** questions regarding our tool as follows, "did you find the visualization tool to be...easy to use?, easy to learn?, fun to use?, efficient to use?, and aesthetically pleasing?".

Participants strongly agreed that our tool was **[multiple choice]** (mean = **[numeric response]**, median = **[numeric response]**), with **[numeric response]** participants selecting "strongly agree" and **[numeric response]** participant selecting "agree".

Participants moderately agreed that our tool was **[multiple choice]** (mean = **[numeric response]**, median = **[numeric response]**), with **[numeric response]** participants selecting "strongly agree", **[numeric response]** participants selecting "agree", and **[numeric response]** participants selecting "neutral".

Participants were neutral about whether our tool was **[multiple choice]** (mean = **[numeric response]**, median = **[numeric response]**), with **[numeric response]** participants selecting "neutral", **[numeric response]** participants selecting "agree", and **[numeric response]** participants selecting "disagree".

Participants were uniformly distributed when asked if our tool was **[multiple choice]** (mean = **[numeric response]**, median = **[numeric response]**), with **[numeric response]** participants selecting "strongly agree", **[numeric response]** participants selecting "agree", **[numeric response]** participant selecting "neutral", and **[numeric response]** participants selecting "disagree".

Participants were split when asked if our tool was **[multiple choice]** (mean = **[numeric response]**, median = **[numeric response]**), with **[numeric response]** participants selecting "strongly agree", **[numeric response]** participants selecting "agree", **[numeric response]** participant selecting "neutral", and **[numeric response]** participants selecting "disagree".

Based on these results, we conclude that our tool **[free response]**.

3.3 Data Processing

In order to empirically analyze the responses they first needed to be scored and processed. We used a summed error metric to score responses, where zero would mean there was no errors in the response. For multiple choice questions, participants received one point for an incorrect answer and no points for an correct answer. For numeric questions, error was calculated according to the following formula: $|response - actual|$. Each distribution had its own set of questions starting with a multiple choice question asking participants to identify the distribution in question. If participants choose incorrectly, their subsequent responses would be incorrect as well since they would be using the wrong distribution. To account for this the numeric questions were scored based on the participants response to the multiple choice question. For example if participants selected "easy to use" instead of "fun to use" their numeric responses would be scored from the "easy to use" data. Once the error metrics for each question was calculated, outliers lesser or greater than 2.5 the standard deviation were removed. Responses with missing data were also removed. These measures were taken to filter out participants who rushed to the end without answering, or by guessing randomly. During this process it came to our attention that we had mistakenly used a different dataset for the bar chart stimuli. This data significantly changed the distribution and answers. Furthermore, we could not locate the data that was used to generate this stimuli, and as a result had to remove the bar chart responses from analysis. Finally, we averaged the error metrics for each participant and distribution resulting in five error metric scores per participant.

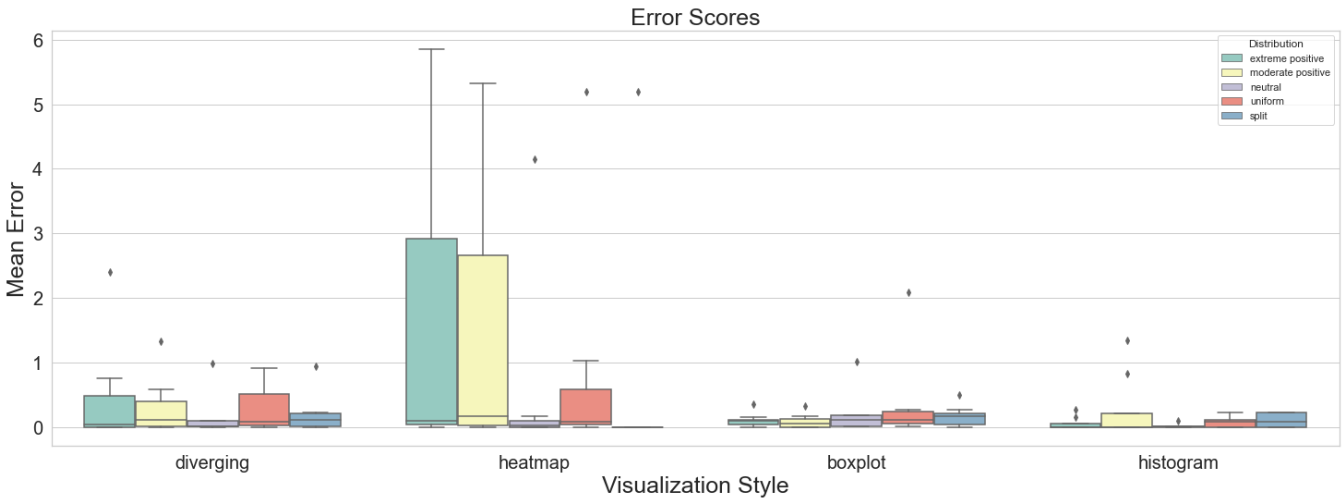

Figure 2: Mean error scores across all participants for each visualization and underlying distribution

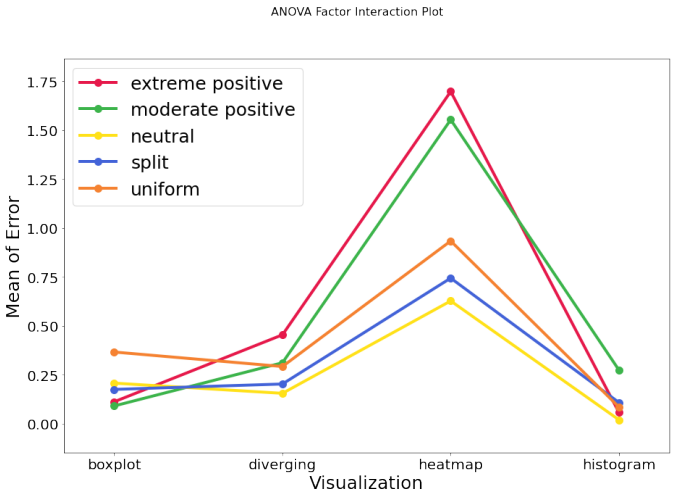

Figure 3: An interaction plot encoding mean error of each visualization and distribution type.

4 Results

We used a two-factor ANOVA to test the research hypotheses. The response variables are error while underlying distribution and visualization style are our covariates, along with an interaction term. We found visualization style to have a statistically significant effect on accuracy ($p = 0.0003$). However, we did not observe any statistically significant results for underlying distribution ($p = 0.7460$) and the interaction of our covariates ($p = 0.9666$). After visualizing the mean error interaction between our factors (Figure: 3) the effects become clear. The boxplot, diverging stacked bar chart, and histogram stimuli all perform relatively the same with small effect sizes between distribution styles. Our heatmap stimuli was the clear outlier with higher error across all five distributions when compared to the other visualization styles. The heatmap performed especially poorly when it came to the extreme positive and moderate positive distributions.

5 Discussion

The study showed the following main findings:

- Heatmaps performed the worst among the visualization styles
- Heatmaps had the most variation in performance between the underlying distributions
- Boxplot, diverging stacked bar chart, and histogram all had close to the same performance.

In this section we will discuss these outcomes, what may have caused them, and what their implications could be.

5.1 Bar Chart of Means Visualization

During the analysis section of the study, we realized our actual values for the bar chart visualization was lost. We were unable to calculate accuracy on our participant responses for this visualization type without the actual values. Many of our visualization styles were created using the same generated data, but we could not make that assumption for the bar chart's data. Due to the limiting nature of this visualization (i.e., no data shown for counts or distribution of counts), participants were forced to make estimations for many of the responses. We are still interested in these estimations and how they compare to study participants' takeaway of the data from the visualization, and will include this visualization in future deployment of this study.

5.2 Free Response

The last question in our study was designed to understand participants' overall takeaways from the trends seen in the Likert data visualization. We allowed an open response format to eliminate bias that comes with limited responses and nuances of the English language. Participants tended towards responses which repeated earlier observations in the distributions of data for each Likert item. For example, many participants responded simply, "easy to use" or "easy to learn" with very few offering additional description on the overall trend such as, "was effective" or "overall neutral tool". It is unclear what participants' takeaways of each visualization were from their responses, redesign of this question or a larger sample size may help provide more insight.

5.3 Sample Expertise

This study is designed to produce recommendations using empirical evidence to help visualization authors choose the best visual representation of their Likert scale data. The target audience of these

visualizations are the visualization authors and readers of visualization papers, and appropriately we classify this group to be experts. The sample of participants from our study includes undergraduate computer science students with a semester of data visualization knowledge. We classify this group of participants as knowledgeable, but not experts. Student results are invaluable as they are often the readers of papers, however it would help to build more confidence in follow-up studies to include expert participants.

5.4 Stimuli Considerations

For all visualizations, the underlying distributions for each Likert item were the same, however the exact counts were not the same. The box plot, histogram, and diverging stacked bar chart visualizations encoded the same data, whereas the barchart as well as the heatmap did not. This can be seen when comparing the "easy to use" likert item counts from the heatmap and the diverging stacked bar chart visualizations; there are 6 responses for "5=strongly agree" on the former but 7 on the latter. The loss of our bar chart response data in our analysis could have been avoided if we used the same generated Likert data across all visualizations.

We constructed each visualization to follow best practices and to encode the data to the best of its ability. Improvements can still be made on these visualizations such as adding titles and making sure the size and aspect ratio of each is normalized for fair comparison.

5.5 Recommendations

We believe that all five visualization styles we tested in this study have a time and a place to be used for Likert style data. While heatmaps had the worst performance, they are still desirable for the space efficiency. The main weakness of heatmaps is its reliance on a color scale to encode the distribution of the data. Previous study's [1] have shown that color is perceptually worse than position, but there are also several ways this could be potentially improved. For example, it is possible that clustering the heatmap and using a diverging color scale may have helped readers identify the trend in the data easier.

We believe boxplot, diverging stacked bar chart, and histogram all performed well because they encoded the data using position and could explicitly show the distribution of the data. Boxplots are capable of encoding the most information showing the mean or median, standard deviation, interquartile range, and using dots to show individual responses. However, this method of showing individual responses does not scale to large responses styles. In this case histograms would be a good option to consider, but it is worth considering that histograms will take up much more space than the other visualization styles.

6 Limitations and Future Work

We encountered several issues with our study design that we can learn from and address in follow up studies. As stated previously, it was possible for participants to answer inaccurately for certain questions which would affect their answers to subsequent questions. We were able to account for this through our response scoring method, but ideally participants would always have the best chance to answer correctly. Furthermore, our chosen survey platform made it difficult to construct the questions in precise manner we wanted. Our goal was to create a fill in the blank style response template, however this was not possible using Qualtrics and we had to break down our questions into a more traditional survey format. We would like to address these points in a follow-up study where we could refine our question template, as well as us a custom web-page as our apparatus to give us more control over the survey style.

7 Conclusion

In this study, we compared the accuracy in participant interpretability of five visual representations of Likert response data. Our findings support visualization style has a significant effect on accuracy. We found the boxplot, diverging stacked bar chart, and histogram all performed relatively the same, and the heatmap performed with higher error across all distributions. While heatmaps performed the worst, we believe each visualization style could be potentially improved and used for Likert data in different situations. More empirical studies are required to start forming a helpful set of guidelines for the visual representation of Likert response data.

References

[1] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[2] H. H. Friedman and T. Amoo. Rating the rating scales. *Friedman, Hershey H. and Amoo, Taiwo (1999)." Rating the Rating Scales." Journal of Marketing Management, Winter*, pp. 114–123, 1999.

[3] S. E. Harpe. How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6):836–850, 2015.

[4] J. Hartley and L. R. Betts. Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on likert-type scales. *International Journal of Social Research Methodology*, 13(1):17–27, 2010.

[5] R. M. Heiberger, N. B. Robbins, et al. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software*, 57(5):1–32, 2014.

[6] T. M. Liddell and J. K. Kruschke. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348, 2018.

[7] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[8] G. Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[9] N. Schwarz, B. Knäuper, H.-J. Hippler, E. Noelle-Neumann, and L. Clark. Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55(4):570–582, 1991.