

Exploring multi-attribute nonprofit survey data

Justin Chin

University of Colorado
Boulder

Bryan Costanza

University of Colorado
Boulder

Bret McSpadden

University of Colorado
Boulder

Ben Niu

University of Colorado
Boulder

ABSTRACT

Although many nonprofits are collecting data, few are analyzing it due to a lack of skill, time, or other factors. Using data from a nonprofit organization that had not been analyzed holistically before, we attempted to better understand the data through data visualization, build a tool that allowed the data to be presented clearly, and give feedback on the dataset to improve future analysis. As a final deliverable, we designed an interactive visualization that shows the correlation between many program factors and volunteer teachers' satisfaction with the program. Through this visualization, we found that the volunteers' ratings of field staff are the most highly correlated factor to satisfaction.

Keywords

Nonprofit, data visualization, Likert-type data, multi-attribute data

1. INTRODUCTION

The collection of data is widespread throughout a variety of organizations but analyzing data presents a unique challenge for small nonprofit organizations [1]. These small nonprofits recognize that the data they collect can help demonstrate impact and help make organizational decisions, but there is often a lack of resources, time, or expertise needed to achieve these ends [2]. Therefore, although a lot of data is being generated and collected, it is rarely being analyzed as a whole [3]. The stakes for not doing so are high in an industry moving toward data-informed models for decision making and development.

For this project, we looked at survey data from an international nonprofit that places volunteer teachers abroad for one-year contracts in partnership with local organizations or government ministries. The volunteers attend three trainings throughout the year: Orientation (OR), Mid-Service (MS), and End-of-Service (EOS). Following these trainings, the volunteers are given an opportunity to provide feedback on elements of the program, including training, support, and safety. The surveys are conducted online via Survey Monkey with Likert-type scales and free-response questions.

As described in the research cited above, the organization collects data, but has not analyzed the data as a whole, instead typically looking at cohorts' responses following the trainings to see individual feedback on staff and any outliers regarding safety or other factors. Although the organization had a general sense of some of the feedback, it was not possible to look more broadly at the data to draw out other insights.

With these challenges in mind, we were interested in looking at how data visualization can help the organization better understand

the survey responses. We were also interested in providing process improvement suggestions for collecting and analyzing data.

2. RELATED WORK

As part of the design process we looked at a variety of resources to better understand Likert-type data and how it is visualized. The foundations of this data type are in behavioral science, but are now used widely to attempt to measure sentiment (and other qualitative metrics) quantitatively [4]. The methodology for analyzing Likert-type data is vast, and not without dissent, but research shows that some criticisms are unfounded [5]. For the purposes of this project, we focused more on visualization techniques, but more robust statistical analysis of the data beyond Pearson's r would be helpful.

In researching visualization of Likert-type data, stacked bar charts, often diverging, appear to be a standard for visualizing individual questions and responses [6]–[8]. Less research was found in comparing Likert-scaled data to other questions, but scatterplots are a common technique when trying to compare two Likert-type questions [9].

3. DESCRIPTION

The goal of our project was to utilize the visualization techniques to analyze and understand a dataset that had not been looked at from a visualization perspective. Our partner organization provided the raw data after removing personally identifiable information. The data included responses from six program countries where the organization places volunteer teachers and represents survey data collected from volunteers at three points during their service: following Orientation, Mid-Service, and End-of-Service trainings. The three surveys are different but included some common questions across the surveys, particularly MS and EOS.

Initial explorations of the data focused on understanding the surveys and pulling out questions that seemed interesting to analyze. Our initial questions influencing our exploration were, "How are different aspects of the programming rated?" and "What contributes to the growth of volunteers?" In answering these questions, we were interested in how volunteer responses changed during their service from beginning to end.

For the purposes of this project, we focused solely on the Likert-type questions and ignored all free-response answers. We first looked through all the surveys and grouped questions by themes: self-assessment, education office, staff, safety, and training. There were also questions relating to satisfaction and whether the volunteer would recommend the program to others. Overall, only one set of questions, the self-assessment, were asked across all three surveys, but items about staff support, safety, and other program features were asked across MS and EOS surveys.

The raw data was somewhat challenging to work with due to the formatting from the Survey Monkey export. After identifying the sets of questions to analyze, we cleaned up the sheets manually to standardize the header rows and processed the data using Python in Jupyter Notebook to calculate the average responses across countries for the three surveys. We combined that data into one CSV file. Although we would have liked to look at individual responses, this was not possible because individual respondents could only be identified by their name or email address which were removed by the organization to protect the respondents' privacy.

After cleaning the data, we used Tableau to better understand the cleaned data and prototype potential visualizations. We explored visualizing the data over time across surveys as well as individual metrics and categories of questions. During these initial explorations it was clear that there were interesting trends and unexpected outcomes in the data. Some examples such as the self-reported ratings trends, low ratings for educational support tools, and the difference between satisfaction and recommendation are included below (Figure 1, Figure 2, Figure 3).

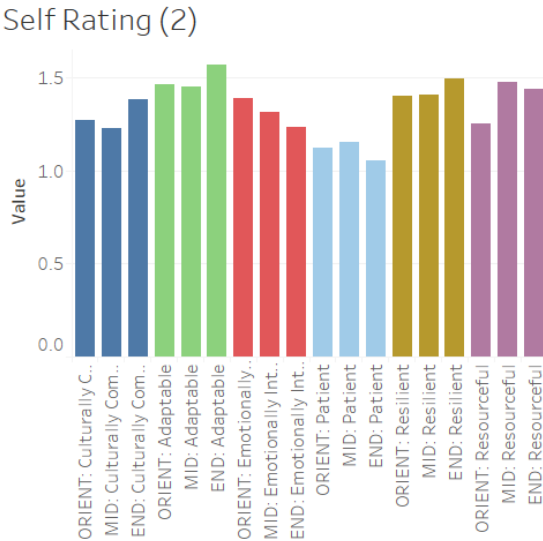


Figure 1: Self rating metrics over time

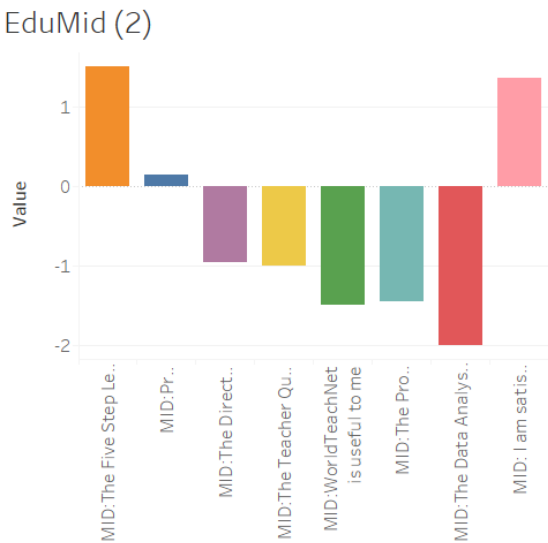


Figure 2: Education support tools, MS

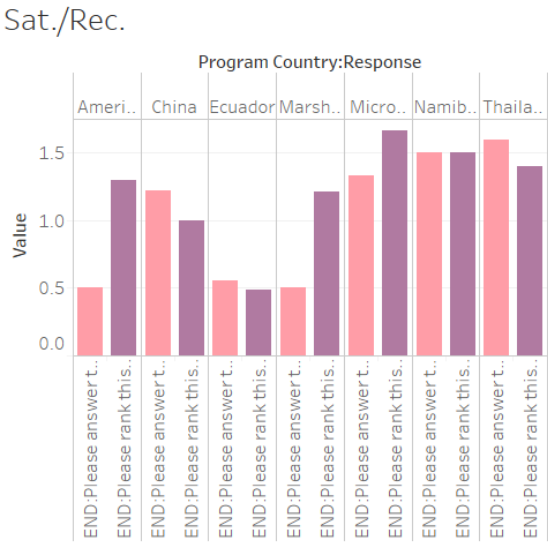


Figure 3: Satisfaction and Recommendation across programs

In prototyping, it was also clear that the number of metrics made it difficult to compare different metrics individually and across countries due to having too much information on the screen in some prototypes. Additionally, although the static bar graphs were interesting, because of the challenges in comparing data, it was hard to further probe the data beyond simple charts. We had a better understanding of the data, but had more questions around how the various questions affected satisfaction and recommendation of the various programs, as well as how the data compared among programs.

To further explore the dataset, we generated a prototype in Tableau comparing averages of individual metrics to the average satisfaction and recommendation ratings. Here it was clear there was some correlation but overdraw and general clutter made it difficult to understand exactly what was going on among all the data (Figure 4).

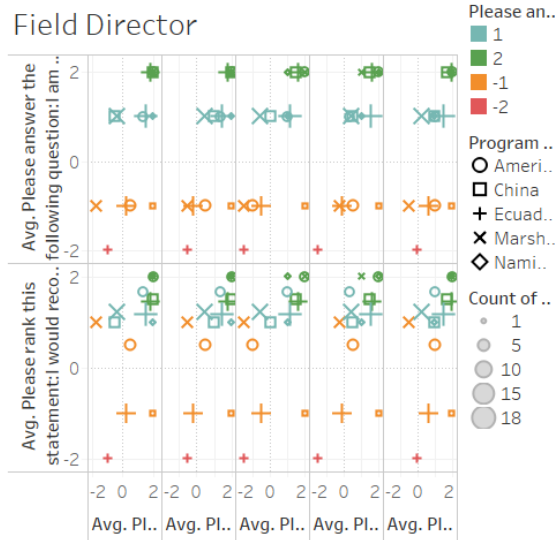


Figure 4: Staff metrics compared to recommend & satisfaction

From this prototype, we generated a new set of questions to answer to help narrow down our final design: “How is WT rated overall? By country?”, “Across programs, what factors are rated highly?”, “What factors are rated poorly?”, “Of the highly rated factors, which ones are mostly correlated with satisfaction?”, “Of the poorly rated factors, which ones are mostly correlated with dissatisfaction?”, “Outliers? Why?”

To accomplish the more complex task of looking at what factors correlate with satisfaction, we designed an interactive visualization to display the r value between questions and satisfaction. We chose satisfaction over recommendation because the two metrics were quite similar, but the satisfaction scores showed a greater numerical spread. Questions were manually divided up into six categories: education tools, impact on self or country, safety, self-reported personality qualities, field staff, and support factors such as a site visit. The visualization was built using the Bokeh library in Jupyter Notebook and put into a HTML webpage. Most of the pre-visualization data analysis was done in Excel. In addition to using online examples for certain tasks and interactions, we used code from project two [10]. Certain programming challenges led to some design choices such as the buttons over a dropdown menu.

The scatterplot on the top left of the page is displayed throughout the visualization and displays each question’s r correlation on the x-axis. The y-axis is the rank, with more highly correlated items (closer to an r value of +1) ranked higher. The y-axis is somewhat confusing because we could not find a way to invert the axis to show “1” as the highest rank at the top, so the rank is not intuitive in how we generally think of it numerically, but we were able to preserve the list in a descending order.

To further explore the correlation, the user can choose a subset of questions and view the top three correlated questions, as well as all staff questions, since those questions were the highest correlated group. The individual graphs showing particular questions are displayed alongside the r value chart and are ordered from highest to lowest correlation. In these charts, countries are coded by color and the number of respondents are coded by circle size. The tooltip displays the charted values. Overdraw was accounted for by displaying multiple tooltips at the same time when points were

overlapping. We chose to only show three of each question, besides staff, since the correlation was generally weak for most categories except staff and support. Safety also only had three questions, so displaying 3 kept the visualizations consistent. This design choice does prevent the user from exploring each question individually.

In addition to exploring the correlation between different questions and volunteer satisfaction, the tool can also be used to generally see how different questions are rated by only looking at the x-axis scale and individual countries. Because the country data is divided by response, this observation is only a generalization. Data from two countries was removed due to lack of respondents.

4. FINDINGS

In exploring this dataset, we had quite a few interesting findings. These can be broken into three categories: from data exploration, from prototyping, and from the final visualization.

From working with the data, a few things stood out. First, although the data was collected in Survey Monkey, the export from the service did not work well with Tableau by default. We also noticed that some of the questions were not scaled the same, with the majority being a 4-point scale, but some with 5-point scales or fewer than 4. Additionally, some questions were not formatted the same, creating unnecessarily messy data that forced added cleaning. Fixing these issues would make further analysis easier, particularly for less-skilled programmers. Additionally, the literature shows that a 5-point Likert-type scale is the norm and helps create equal distance between the ratings. We expressed this in the data by using -1, -0.5, 0.5, and 1 to substitute for Strongly Disagree, Disagree, Agree, and Strongly Agree.

Two other major items stood out about the dataset itself. First, because the data is collected as individual responses across the three surveys, the only individually identifiable information is a respondent’s name or email, but this information was removed at the organizations request to protect the privacy of individual responses. Although we ended up not comparing data across surveys for the final visualization, in our explorations we were forced to aggregate the data. This aggregation and lack of identifiable individual respondents made it challenging to drill down into interesting patterns during the prototyping and added to uncertainty knowing that different respondents may be answering different surveys. Knowing who will process the data, and considering privacy concerns in advance, may help allow for deeper analysis of the data in the future.

Second, the survey contained a lot of open ended responses and, although these may be helpful in better understanding the data, we decided to not attempt to work with those questions, leaving out a lot of valuable information. Better capturing and mapping those responses may be helpful in further analyzing the dataset.

During prototyping it was interesting to be able to quickly look at the data visually. During these explorations it was useful to quickly show unexpected aspects of the data which helped us question our assumptions and identify interesting aspects of the data to explore further or to discard areas of investigation when nothing interesting seemed apparent in initial explorations. During this process, we saw that satisfaction and recommendation did not always get rated the same (Figure 3). The self-reported data measuring resilience, patience, etc. did not generally increase over time, which was surprising (Figure 1). We also noticed that the education support

tools were rated significantly lower than any other factor (Figure 2).

Finally, in the designing and building of the final visualization, we were able to show that staff and support training are the two aspects of the organization's programming that mostly correlate with satisfaction. Staff support factors make up eight of the top ten best predictors of volunteers' satisfaction with the program. The orientation training rating was the most correlated to satisfaction in the area of support training. Similarly, providing a worthwhile experience is the most correlated impact factor, and the five-step lesson plan is the most correlated education support factor. Other items were not very correlated with satisfaction.

In looking specifically at the highest correlated questions, those relating to field staff, it is interesting to see that administrative elements are less correlated than emotional and general support.

In exploring the individual questions, one can see that the highly correlated items follow the expected linear trend of poor rating to poor satisfaction or high rating to high satisfaction. One also can see that many items are generally rated high, regardless of correlation, except for the education items, which are generally rated poorly.

From these observations, one can infer that the organization should focus on hiring good staff members and ensuring its trainings, particularly Orientation, are well organized as these factors are the most correlated to satisfaction. It is also interesting to note that it seems that emotional support is more correlated than administrative support. Therefore, the organization should focus on supporting its staff in supporting volunteers through their service.

REFERENCES

- [1] S. Lohr, "Opinion | Big Data's Impact in the World," *The New York Times*, 11-Feb-2012.
- [2] C. Bopp, E. Harmon, and A. Volda, "Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 3608–3619.
- [3] N. L. Maxwell, D. Rotz, and C. Garcia, "Data and Decision Making: Same Organization, Different Perceptions; Different Organizations, Different Perceptions," *Am. J. Eval.*, vol. 37, no. 4, pp. 463–485, Dec. 2016.
- [4] H. N. B. Jr and D. A. Boone, "Analyzing Likert Data," *J. Ext.*, vol. 50, no. 2, Apr. 2012.
- [5] G. Norman, "Likert scales, levels of measurement and the 'laws' of statistics," *Adv. Health Sci. Educ.*, vol. 15, no. 5, pp. 625–632, Dec. 2010.
- [6] R. Heiberger and N. Robbins, "Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications | Heiberger | Journal of Statistical Software."
- [7] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual Analysis of Multi-Attribute Rankings," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2277–2286, Dec. 2013.
- [8] S. Wexler, "How to Visualize Sentiment and Inclination," *Tableau Software*. [Online]. Available: <https://www.tableau.com/about/blog/2016/1/how-visualize-sentiment-and-inclination-48534>. [Accessed: 03-May-2018].
- [9] swexler, "Likert vs. Likert on a Scatterplot," *Data Revelations*, 16-Sep-2014. [Online]. Available: <http://www.datarevelations.com/likert-vs-likert-on-a-scatterplot.html>. [Accessed: 03-May-2018].
- [10] *project-2-ncwit-team-13 created by GitHub Classroom*. INFO 4602/5602: Information Visualization, 2018.