Perceptrics: A Perceptual Metric for Evaluating Graph Effectiveness

Sukhi Gulati
Stanford University
sgulati3@stanford.edu

Raymond Luong
Stanford University
rayluong@stanford.edu

Julie Ni
Stanford University
jni221@stanford.edu

ABSTRACT
Visualization creators currently lack an easy way to detect how effective graphs are at communicating data and how accessible the graphs are for the visually impaired. We designed a system that evaluates and rates a graph in four primary categories - overall scaling, data-ink proportion, data dimensions, and color contrast ratio - based on common design principles and guidelines that are backed by cognitive science and studies on human perception. Our methods for developing each metric and techniques for calculating the evaluation results are derived from such design principles and confirmed through extensive user testing. The developed software, Perceptrics, offers actionable suggestions for improvement based on which metrics the graph fails to pass, and users can realize those changes in real time on the uploaded graph. In the end, being able not only to identify which aspect of the graph was improvable but also to apply those improvements instantly and see the reflected change in comparison to the original graph was the core feature that facilitated general understanding about cognitive effectiveness. This paper examines the use of an objective, perceptual metric to help visualization creators develop more effective graphs.

Author Keywords
Graphical perception, visualization, chart understanding, graph design, perceptual effectiveness

ACM Classification Keywords
H5.2 [Information Interfaces and Presentation]: User Interfaces – Graphical User Interfaces.

INTRODUCTION
There are well researched and objective design principles, which draw upon rules of human perception human perception, that make graphs more effective and interpretable. However, this research is decentralized across many publications, researchers, and disciplines. Many designers creating visualizations do not have extensive prior knowledge in the field. Even if designers do read about these principles, they are often presented as loose heuristics and not clearly actionable analyses. To make the most of these principles, it is important that they are aggregated and presented in a straightforward fashion. As is, visualization creators lack an easy way to receive feedback on whether their graphs are perceptually effective.

Unlike the subjectivity of aesthetic design, a metric of perceptual effectiveness would provide objective graphical feedback. Graph designers should take perceptual effectiveness into account not just to create clearer graphs in general, but also to ensure that graphs are more easily accessible to differently abled populations. Sufficient color contrast, for example, is important for individuals with visual impairments. A tool that measures perceptual effectiveness is would make the visualization process more approachable to those without background in the field of cognition and perception. Although there are many prolific tools to create visualizations from large datasets, there are few tools intended to refine those visualizations on a basis other than extrapolating trends from the data. For this reason, our goal was to aggregate disparate heuristics of perceptual effectiveness into one single interface or metric.

RELATED WORK
The foundation of our approach lies within the works of experts in the field of graphical perception and information display. These works are “Graphical Perception and Graphical Methods for Analyzing Scientific Data” by William S. Cleveland and Robert McGill [1], and “Visual Display of Quantitative Information” by Edward R. Tufte [2]. Cleveland and McGill uncover elementary tasks for the graphical perception of quantitative information and use these tasks to further our understanding perceptual

effectiveness. Their research was performed in the 1980s and has important implications for future works regarding the display of data. Tufte details how to create effective visualizations and outlines important principles, such as data ink, chart scaling, dimensions of data, and chartjunk, that ultimately affect how a user perceives and interacts with a graph. Our interface draws heavily upon these works to determine what constitutes perceptual effectiveness. We also looked at applications and interfaces similar to what we wanted to create.

From our foundational works, we looked at related applications in the field of effective visualization. Most notable of these works are "ReVision: Automated Classification, Analysis and Redesign of Chart Images" by Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, Jeffrey Heer [3], and "Measuring effectiveness of graph visualizations: A cognitive load perspective" by Weidong Huang, Peter Eades and Seok-Hee Hong [4]. ReVision presents a tool that extracts data and information from a provided graphic and generates a new visualization from the underlying data. Our project starts at the same location – analyzing existing graphics – but we differ in our approach and end goal. ReVision uses an approach to data extraction that is heavily backed by computer vision. Our focus is not to extract data from existing graphs, but to critique those graphs and give designers the tools and suggestions that can aid in their redesign. "Measuring effectiveness of graph visualizations" discusses the cognitive load model in relation to visualizations and how this model is affected by visual, data, and task complexity. An important takeaway from this paper is that by reducing visual complexity through following a standard set of perceptual and design principles, we can offer the user more cognitive resources to allocate to data complexity and task complexity.

METHODS

Our methods can be broken down into two main parts: developing the metric and building the software.

Metric Development

Initially, we worked towards creating one aggregate score that conveyed the overall "perceptual effectiveness" of a graph. After continued thought, discussion, and testing we decided to focus less on creating one effective measure and more on creating an interface to automate feedback based on a few of these measures. Based on the foundational works described above, we selected multiple heuristics to explore. These included: Graph Dimensions, Typography Proportions, Data Point Presentation, Effectiveness of Annotations, Effectiveness of Captions and Titles, Color Scheme, Correctly Mapping Ordinal and Nominal Values to Visual Classifiers, and Matching Data Dimensions to Visual Dimensions. From here we used a few approaches, including testing with volunteers and further research, to examine how these factors interacted with each other and what would be best to evaluate.

To explore the relative weighting of these metrics, we created multiple formulas to aggregate them. The most basic of such formulas was an equal weighting. Given n factors, each factor would be given a decimal score f within 1 - 10. The equally weighted equation was given by

$\sum_{i=0}^{n} \frac{1}{n} f_i$, presented as a percentage of n. We adjusted the relative weightings of each factor f to come up with multiple formulas. To test these formulas, we wanted to see how they compared to human derived scores. We gave volunteers a structured form that walked them through rating graphs. The form asked questions about each graph component that our metrics touched upon (ie: "Are captions effectively written? Do they concisely convey the message of the graph without extraneous information?"). We had volunteers rate a selection of graphs, and we rated those same graphs according to different versions of our formula. After comparing the resultant ratings, there highly variable differences between the metric-derived scores and the human-derived scores. Our metrics were, however, touching upon the qualitative comments of human volunteers. For example, one volunteer noted that the gridlines on a graph were distracting and the same graph lost points in our formulas for those gridlines. From this we realized that the constitutive components of our metrics were often valid, but the arbitrary nature of weighting them in order to create an aggregate score actually took away from their valuable insights. We also did not find prior research that attempted to relatively weight the importance of design principles or principles of perception. After receiving this information from our initial volunteer testing, we decided to focus on creating an interface that could evaluate graphs based on several different measures of perceptual effectiveness. In short, we decided to collect the perceptual metrics into one location, but keep the metrics disaggregated and present them as separate scores.

Here, our first task was to narrow down the list of metrics we would to consider for our interface. We determined the metrics we chose should be easy to automate, backed by perceptual research, and qualitatively significant. We looked over the open-ended comments on graphs from our volunteers and revisited initial research to pick the following four measures of perceptual effectiveness:

1. Overall Scaling: A combination of graph dimensions and typographic proportions.
2. Data Ink Proportion: A heuristic intended to emphasize data and reduce extraneous annotation.
3. Color Contrast Ratio: A measure of the relative luminance of two colors intended to ensure readability.
4. Appropriate Data Dimensions: A binary check that the dimensions of the data match its representation.

Some metrics, although interesting and important, were not objective enough to easily automate. We explored using tonal analysis APIs to evaluate caption effectiveness, for example, but ultimately decided that using such a highly discretionary principle would defeat the purpose of a tool. Formalizing caption effectiveness is an endeavor in and of itself; our tool intends to be more of a proof of concept.

Software Development

After selecting our metrics, we began developing software to assess graphs uploaded by users. We developed our interface using HTML, CSS, and JavaScript. We named our interface "Perceptrics". Below, we will describe the main technical decisions we made when implementing our software including how each metric is calculated:

Parsing Graphs

When deciding the input format for graphs in Perceptrics, we wanted to avoid focusing our problem space in computer vision. For this reason, we do not parse bitmap images. Since many bitmap image processors first render the bitmap to an SVG format, we decided to explore SVGs as a possibility. All user input is uploaded in an SVG format. We use a combination of specified HTML tags and class names to isolate graph components. For example, to analyze a graph title, we expect the input SVG to have an element with the class "title". We use a JavaScript FileReader object to allow users to upload SVG files. Perceptrics in its current state only analyzes bar graphs. However, this is easily extensible because of the structured input format.

Overall Scaling

For overall scaling, we analyze two factors. The first is how closely the graph's dimensions adhere to the golden ratio, as described by Edward Tufte [2]. The second looks at the relative scaling of text on the graph. An accepted heuristic is that the title of the graph should be about twice the size of axis/tick labels on the graph.

Data Ink Proportion

Data ink, a term Tufte uses, is the ink on the graph that displays non-redundant data information. On a digital tool, this ink is translated to pixels. To measure the data ink ratio, we extract both data and non-data tags on the page and use the following formula from Tufte:

Data ink ratio = data ink / total ink used to create graphic

Data elements are marked with the class "data" while non-data elements are the remainder of the elements, such as axes and gridlines. We approach data ink as a heuristic as opposed to a score. There is no specific, ideal data ink proportion. Instead, users should be aware of any extraneous elements and recognize that graphs should seek to minimize extraneous elements and maximize pixels relevant to data presentation. We believe that presenting the ratio and suggesting muting the appearance of certain non-data elements conveys this message.

Color Contrast Ratio

The World Wide Web Consortium (W3C) declared a standard for color contrast in order to promote accessibility. Perceptrics uses this color contrast standard and scores graphs against it. We take every pair of types of elements on the graph that appear adjacent to each other - such as data and gridline, title and background, data and background - and extract the rendered colors to ensure that the contrast meets the standard. The color contrast ratio formula is defined by the W3C and involves first calculating the relative luminance of colors and then a ratio between their relative luminance scores. Relative luminance is calculated using the linear RGB values of a color and weighting the individual values according to constants that reflect the luminosity function.

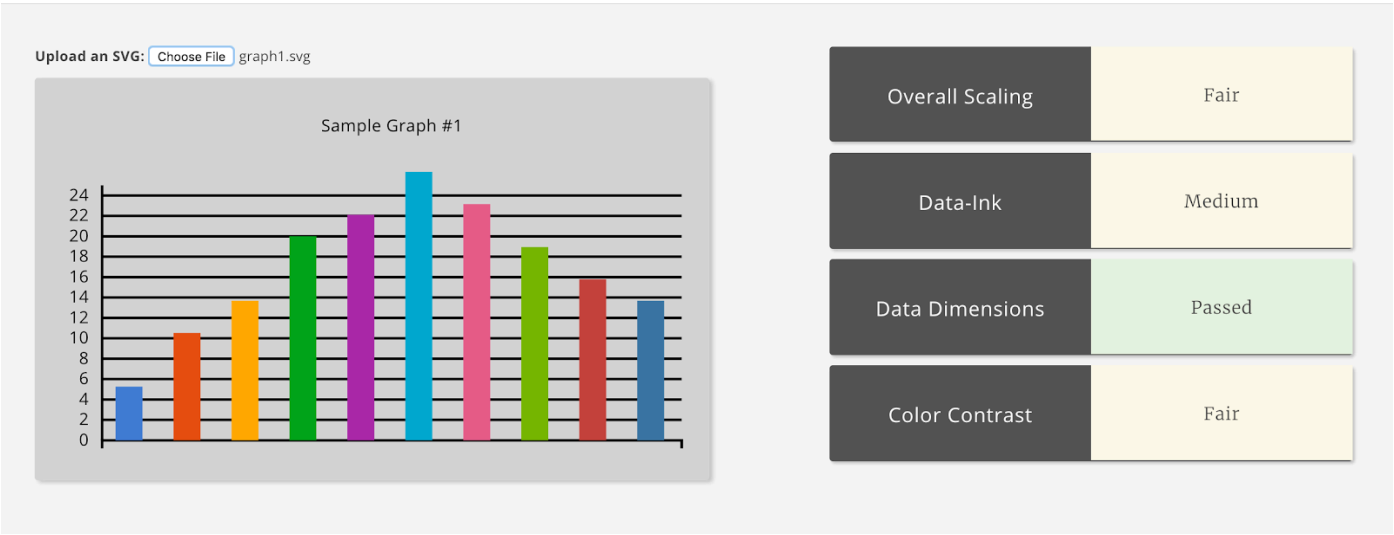

Figure 1: Upon initial upload, Perceptrics automatically rates the SVG graph.

Appropriate Data Dimension

Bar graphs are most appropriate to visualize a one-dimensional measure because the bars grow in only length. Thus, bar graphs should visualize a nominal measure (categories) against a non-nominal measure. We constrain the input to have a "nominal" or "non-nominal" class on the axes and check that there is exactly one nominal and one non-nominal dimension. This metric is binary in nature, rather than a variable score like the other three metrics.

RESULTS

Here, we will give an overview of the Perceptrics software in its current state. Perceptrics is a web application best suited for Google Chrome. Users upload graphs in an SVG format. When a graph is uploaded, Perceptrics runs its algorithm and displays the corresponding assessment for each metric. Users have the option of clicking a metric to drill down and learn more about the metric as well as why they received that score. For some metrics, Perceptrics also offers potential actionable items that users can take to improve the graph's score. Selecting these actionable items renders the change in real time on the uploaded graph to help users learn. Our goal is not for the changes we render to be the final output, but to give users an idea of how they may iterate upon their graph.

When a graph with a poor typography ratio fails the scaling metric, for example, users may choose to enlarge the graph title so that it is twice as large as the axis labels for maximum differentiation.

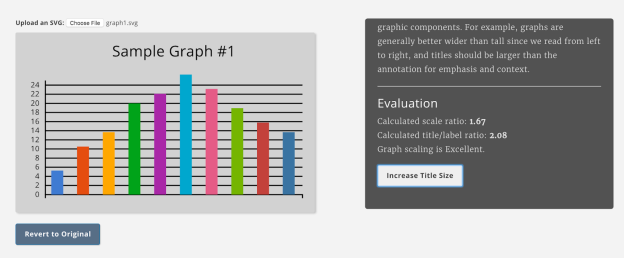

Figure 2: Sample Graph #1 after user increases title size.

When a graph exhibits a low data-ink proportion, some users chose to remove gridlines completely, while other people chose to make them thinner. Many stated that they appreciated having two options because while removing them would eliminate non-data-related elements, gridlines could be helpful in providing a reference point for reading quantitative measures.

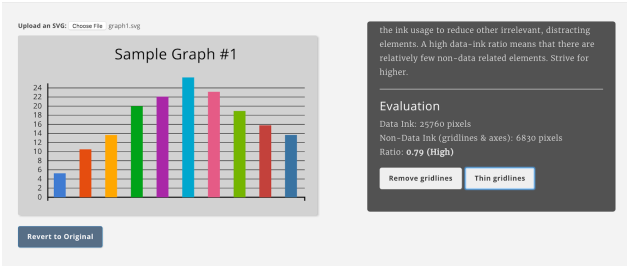

Figure 3: Sample Graph #1 after user thins the gridlines.

When a graph did not possess enough color contrast, users are presented the option to either lighten the background or darken the data points to see how much contrast was necessary to support accessibility for the visually impaired.

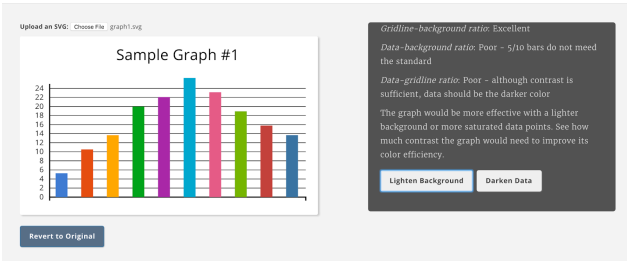

Figure 4: Sample Graph #1 after lightening the background.

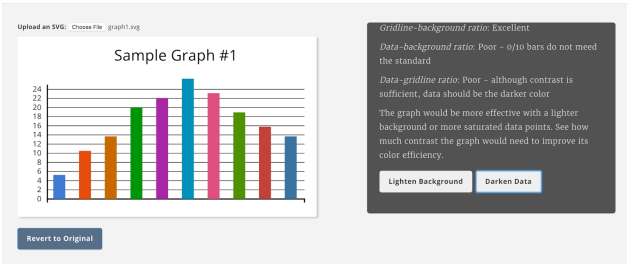

Figure 5: Sample Graph #1 after darkening the data in addition to lightening the background.

After each modification, the graph's evaluated metrics are updated in real time to reflect the improvement.

Rating the graph takes a few milliseconds and the evaluation rates show up instantly on the web interface. We found that Perceptrics' actionable items were necessary to guarantee that users understood the metrics and how those features would enhance perceptual effectiveness. Most users did not show any audible signs of comprehension or exhibit "A-ha" moments until they clicked on the suggestion action item and saw the corresponding change reflected on the graph. We observed users using our application before we added in the dynamic graph updated and users largely complained that they still might be unsure how to improve their graphs. Adding the option to

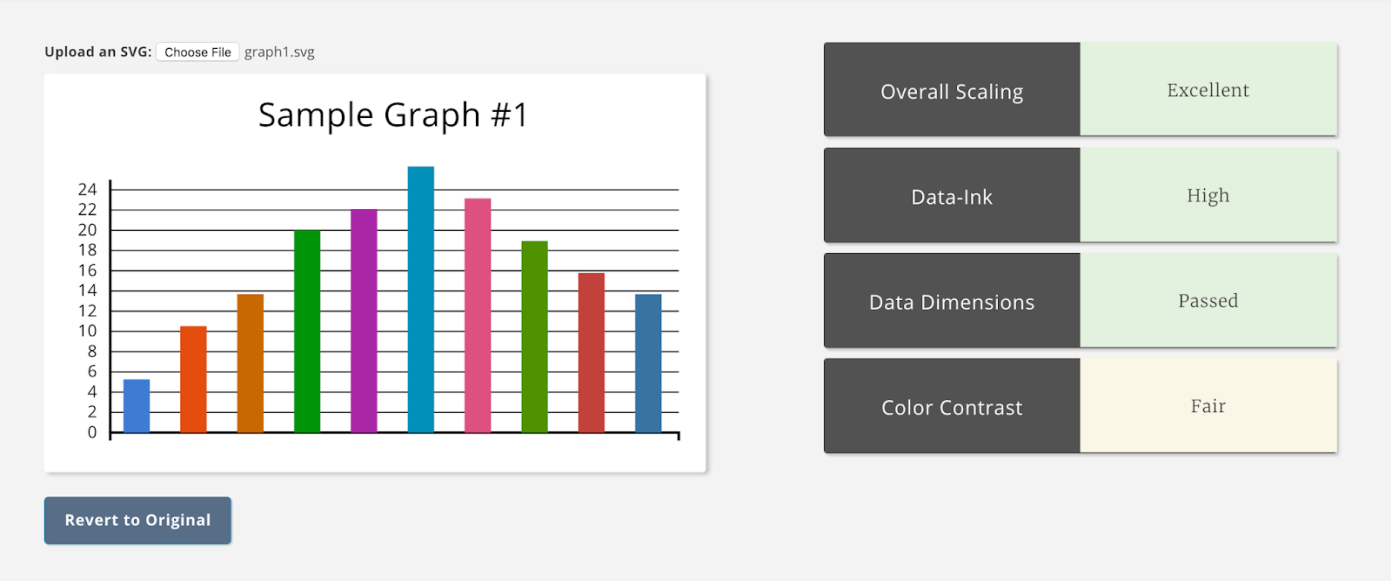

Figure 6: Sample Graph #1 after improving upon some metrics.

demonstrate example changes created a notable difference in perceived usefulness of our tool. This indicated that real time graph manipulation is crucial in facilitating not only understanding but future application of these same techniques. Overall, most users felt that the graph rendered with all applied suggested improvements were markedly better than the original graphs. It is particularly striking that these same users - except for one - deemed the initial, unimproved graph adequate and recognized its flaws only in comparison to the redesigned version.

In addition to the positive feedback, users were curious about whether Perceptrics would be more of a one time tool to teach how to create an effective visualization or whether it would be a continuous tool used to iterate on the same graph to make it better. It is also worth noting that although no users felt negatively about graph changes, some users felt indifferent towards changes in the graph and would hope for something more substantial.

DISCUSSION

Perceptrics proves to be an effective proof of concept for graphic analysis tools based on measures of perceptual effectiveness. In the status quo, many otherwise accessible design principles remain functionally inaccessible for designers who are not always well versed in the literature on data visualization. The goal of Perceptrics was to provide an easy way to learn about and implement visualization features that are perceptually effective.

Initial user testing suggests that a few things are important in accomplishing this goal. First, users largely benefit from example. Perceptrics' attempt to redesign an existing graph - as opposed to create a new one - was quite warmly received. This was a large change in Perceptrics' as compared to the previous work that inspired it. ReVision, the graph redesigner that partially inspired our work, was focused on creating good representations from extracted data. However, the act of iterating upon the original graph had a large impact on our users that would likely not be captured by the process of making an entirely new graph. It is worth noting that in the case of our user testing, users did not create the graphs we tested. Instead, we showed them sample graphs that we created and allowed them to use our tool to edit those graphs. It is possible, and worth mentioning, that designers who used the tool on their own graphs may have more varied reactions. Users we tested on did not have attachments to the graph being edited. Whether reactions are positive or negative, though, showing the suggested graph alterations in real time is certainly an effective way to convey exactly what is meant by a given metric. Because perceptual metrics must be precise, and consequently hard to imagine with accuracy, tools that build upon our work should certainly place emphasis on the dynamic visualization of perceptual metrics.

Second, some users that recognized the effectiveness of the improved graphs initially deemed the original graphs as adequate. While our users were not designers, they all had made graphs before for various endeavors and believed that they would make graphs in the future. This shows, that among our users at least, perceptual effectiveness is not particularly intuitive. Users were quick to have an opinion on whether a graph was aesthetically pleasing. They might, for example, dislike the color scheme or font. However, users were less hesitant to call a graph "effective". This notion underscores an important aspect of our endeavor: the distinction between understandability and perceptual effectiveness. If individuals can discern the meaning behind a graph without much effort, they might consider that graph effective. However, individuals have highly variable ability. Perceptual effectiveness seeks to minimize the unconscious processing power that goes into understanding a graph. Enforcing principles of perceptual effectiveness seeks to

increase the probably that any given individual would have an easier time parsing through components of the graph than if those principles were not enforced. Thus, perceptual effectiveness is a necessary but not sufficient component of understandability. Because a graph can be understandable but not perceptually effective, this may contribute to users' tendency to deem perceptually ineffective graphs as overall effective. This also proves the need for a tool like Perceptrics. If it is indeed difficult for individuals to develop an intuition for perceptual effectiveness, then tools which enforce principles like precise scaling may be beneficial to experienced visualization creators as well as novice ones. Of course, our conjectures regarding the intuition behind perceptual effectiveness are hypotheses based on a very small sample size and largely qualitative observation and discussion. We wanted to highlight, however, this observation from our user testing because of its potential to be a site of further research and the degree to which it underscores the potential usefulness of Perceptrics.

In short, we believe Perceptrics and user reactions to the tool prove both the need and the capability to aggregate the currently disparate research on perceptual effectiveness and present it in an accessible fashion. Perceptrics is based on equations about scaling and color that many people creating graphic visualizations will likely not be familiar with. The ease with which users were able to apply these principles to graphs using Perceptrics demonstrates that we have the ability to make these metrics a new precedent in data visualization - at least in simple graphical systems.

FUTURE WORK

A potential avenue for future research involves loosening constraints on user input. In its current state, Perceptrics requires a very specific user input in the form of an SVG file with specific class and tag names. In addition, Perceptrics currently only processes bar graphs. Loosening constraints on input graphs would enable users to analyze a wider variety of graph types, such as line graphs, pie charts, scatter plots, and box-and-whisker plots. Analyzing graphs based on tags makes it easy to extend the accepted graph formats if we do not loosen the constrained nature of user input. For example, calculating data ink assumes that elements with the "data" tag have a rectangular area. Adding minimal additional tags to the user input to classify the representation of data points or simply using existing SVG classifications would go a long way in extending the input capabilities of Perceptrics. Therefore, this would be a good short term goal.

Allowing users to upload bitmap images would also further loosen the constraints on user input. However, this presents an interesting computer vision problem. Our research team is not well versed in the field of computer vision, but believes that the ability to analyze bitmap images would help extend the reach of a tool like Perceptrics. Additionally, adding bitmap images would help designers to evaluate their graphs in earlier stages of the iteration process (such as sketching).

Another avenue for future work is adding more nuance to the metric calculations. Our metric calculations are currently simple, in part because we scoped our project to only a narrow range of acceptable inputs and possible graph formats. Regarding the overall scaling, our metric is highly dependent on the golden ratio as described by Edward Tufte. Two other methods of computing an aspect ratio that we have yet to explore are Cleveland's median-absolute-slope method and optimization technique of banking to 45. Additionally, it would be ideal to consider the shape and nature of the data set when suggesting rescaling options.

Regarding the data ink calculation, our algorithm considers non-data ink to be axes and gridlines. If we were to loosen the constraints on user input and allow for a wider variety of graphs, our definition of non-data ink would consequently have to be extended to accommodate different styles of graphs. The mentality behind the data ink metric is to reduce extraneous information. Therefore, recognizing annotations and labels as extraneous and evaluating their necessity would greatly refine this metric.

Regarding the color contrast metric, our algorithm currently only processes solid color data bars. This can be extended to analyze non-solid colors such as gradients and patterns, as well as to consider the relative harmony of color-encoded data. In addition, our color contrast score assigns an equal weighting to the color contrast between every pair of elements, an assumption that we took for the sake of simplicity but has room for further research and exploration.

Perceptrics at its core is a tool for measuring graphical effectiveness. Offering graphical suggestions was not the primary focus of our work, though we felt that it had a place in Perceptrics because it furthers the user's learning process and offers actionable items that the user can easily visualize. There is much room left to explore in the graphical suggestions we provide and perhaps we can consider how a tool such as ReVision [2] redesigns graphs. Finally, further user testing would refine our metrics and perhaps reopen the possibility of one aggregate score.

CONCLUSION

Here we have presented Perceptrics, an application which accepts user created graphs before evaluating and then iterating upon those graphs based on perceptual design principles. Suggesting relatively low-effort redesigns to graphs based on objective principles has great potential to set a precedent of accessible and effective graph design. Perceptrics can also be a valuable educational tool for new designers and a sanity check for experienced visualization designers. Overall, we are excited to see where Perceptrics and new hypotheses born in its development will go.

ACKNOWLEDGMENTS

We would like to thank the teaching staff of CS448B at Stanford University, namely Professor Maneesh Agrawala and Course Assistants Scott Cheng, Ludwig Schubert, and Peter Washington.

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. E. R. Tufte. The Visual Display of Quantitative Information. Graphics Press, 1983.
3. M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. ReVision: Automated Classification, Analysis and Redesign of Chart Images. In UIST 2011, pages 393-402.
4. W. Huang, P. Eades and S. Hong. Measuring effectiveness of graph visualizations: A cognitive load perspective. Information Visualization Vol. 8, 3, 139 – 152.