Choropleth Maps Can Convey Magnitude Through the Range of the Accompanying Color Legend

Duncan Bradley   Boshuo Zhang   Caroline Jay
Andrew J. Stewart

Data visualization software provides the ability to create highly customizable choropleth maps. This presents an abundance of design choices. The color legend, one particular aspect of choropleth map design, has the potential to effectively convey data points' magnitudes (how large or small they are). Color legends present the mapping between a specific range of colors and a specific range of numerical values. In this experiment, we demonstrate that manipulating this range affects interpretations of the magnitude of plotted values. Participants (N = 100) judged the urgency of addressing pollution levels as greater when the color legend's upper bound was equal to the maximum plotted value, compared to when it was significantly larger than the maximum plotted value. This provides insight into the cognitive processing of plotted data in choropleth maps that are designed to promote inferences about overall magnitude.

Introduction

To make sense of statistics presented in newspaper articles or scientific reports, it is often important to interpret their meaning in context. This may involve determining whether the presented values represent large or small numbers. Data visualizations are often used to convey statistics, so understanding how these tools may communicate data points' magnitudes is crucial.

Choropleth maps employ colors to represent values and are typically used to convey spatial variability. In order to aid discrimination and facilitate identification of spatial patterns, values are often encoded using the entire range of the chosen color palette. Thus, the range of values on the accompanying color legend typically consists of only those values which were observed. However, this is not the only application for a choropleth map. In certain cases, displaying values' *absolute* magnitudes may be considered more pertinent than displaying their *relative* magnitudes. This would allow a viewer to gauge, on the whole, how large or small presented values are, in context. To communicate this, the range of values on the accompanying color legend may include values which were not observed but remain relevant nonetheless. Designers may wish to sacrifice discrimination ability for an overt display of magnitude, in order to convey their intended message.

Indeed, choropleth maps displaying overall magnitudes have been used in practice. Figure 1 depicts data concerning public support for a federal ban on abortion in the U.S. The accompanying color legend presents the entire range of possible values: from 0% to 100% support. Since plotted values do not exceed 30%, their magnitudes appear small, in context. In addition, whereas a typical color scale would amplify differences between regions, this design presents variability between states as low. This lends credibility to the notion that, for this aspect of a divisive issue, public support is consistently low across the U.S.

This paper explores cognitive processing of overall magnitude in choropleth maps. Through an empirical study, we demonstrate that color legends, which depict the mapping between colors and numerical values, can imply whether plotted values are large or small. Even when the mapping between color and numerical value remains the same, the range of the color legend provides a crucial source of context. The relationship between this range and the plotted data influences viewers' interpretations of magnitude.

Related Work

Communicating Magnitude Through Data Visualization

Empirical studies in various scientific fields have explored how interpretations of magnitude are influenced by data visualization design choices.

Recently, the practice of y-axis truncation has enjoyed attention in experiments at the intersection of the disciplines of data visualization and psychology. Y-axis truncation refers to the practice of minimizing the range of values that appear on the y-axis. This typically involves starting the y-axis at a value greater than zero (Correll *et al.* 2020). However, some experiments on y-axis truncation have employed axes that are roughly symmetrical about the plotted data (Witt 2019). Truncation effects are therefore not just associated with the exclusion of a zero value, but also the exclusion of values *above* the observed data, which make differences appear smaller. Thus, more generally, truncation effects illustrate people's treatment of axes as implicit scales for making qualitative judgements about presented data.

Research on the effects of y-axis truncation has focused on how this practice can alter people's interpretations of the magnitude of the difference between plotted values. Demonstrating the effect of y-axis truncation with a large online sample, Pandey *et al.* (2015) found that

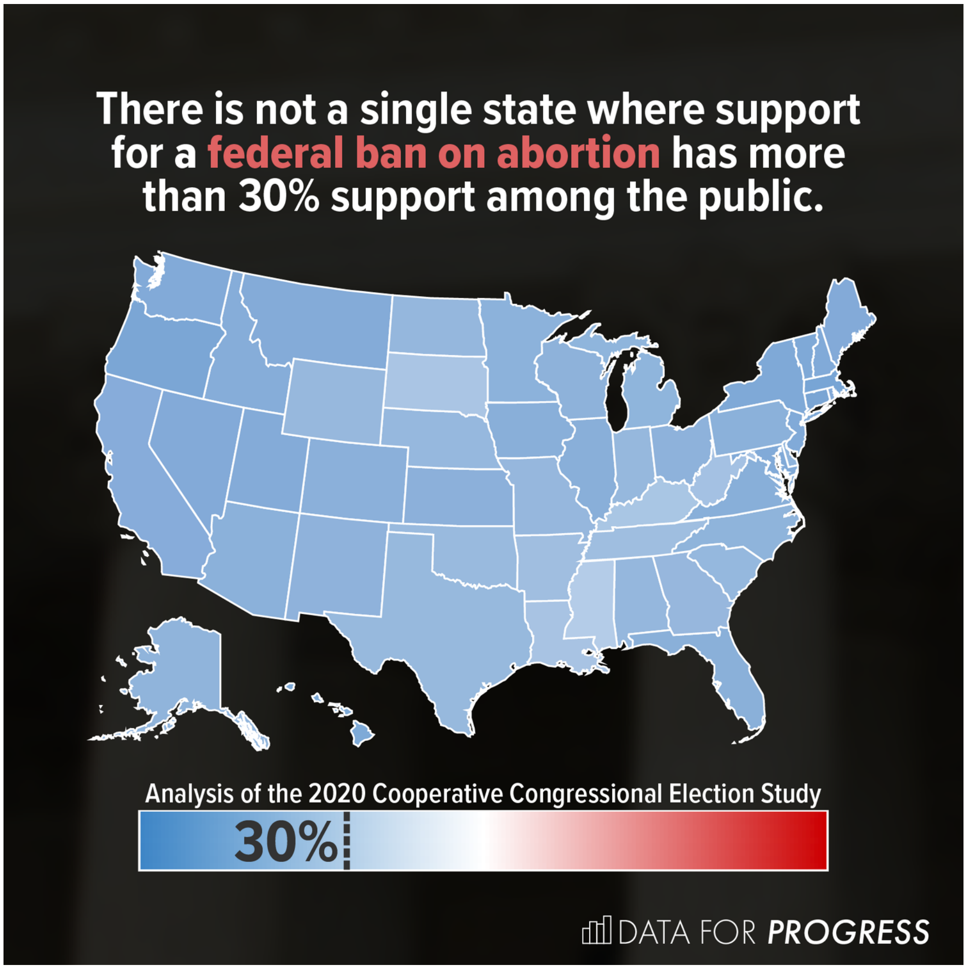
There is not a single state where support for a federal ban on abortion has more than 30% support among the public.
Analysis of the 2020 Cooperative Congressional Election Study
30%
DATA FOR PROGRESS

Figure 1: A choropleth map displaying data from an analysis of state-level public support for a federal ban on abortion in the U.S (Fischer and Ali 2021). The color legend employs a diverging blue-red color palette, with white in the center, showing the full range of possible values. The 30% point is marked with a dotted line and labeled to indicate that no state exceeds this level of support. Reproduced with permission.

ratings of the magnitude of the difference between values were greater when a truncated axis was used to display the difference between safe drinking water levels in two towns. In both bar charts and line charts, increasing the degree of truncation produces increasing estimations of the severity of the difference between values (Correll *et al.* 2020). Encouraging careful attention to plotted data by ensuring that numerical values are read precisely does not eliminate this effect (Correll *et al.* 2020). Warnings somewhat reduce, but do not eradicate, the difference between interpretations of truncated and non-truncated charts (Yang *et al.* 2021). Visual indicators of truncation are also ineffective (Correll *et al.* 2020). Driessen *et al.* (2022) observed smaller effects, but were concerned with numerical estimates of differences between values, rather than subjective interpretations.

Witt (2019) demonstrated that using the widest possible y-axis range diminishes a viewer's sensitivity, which is the ability to distinguish between different degrees of separation between values. On the other hand, using the smallest possible y-axis range increases bias in interpretation (i.e., the extent to which judgments of the magnitude of difference deviate from actual effect sizes). To maximize sensitivity and minimize bias, and to ensure correspondence between the appearance of the difference and the reality, Witt suggests using a range of 1-2 standard deviations for y-axis limits.

Witt's (Witt 2019) recommendations are prescribed for disciplines which use standardized effect sizes (e.g., Cohen's d) in the reporting of data and statistics. Correll *et al.* (2020) provide more general advice relevant to those in all disciplines: the appearance of differences in a visualization should be appropriate for the specific data. Therefore the decision whether or not to truncate an axis depends on the real-world magnitude of the difference, and ultimately designers should ensure they represent this faithfully. Evidence suggests that viewers interpret the axis range as a representation of the relevant numerical context within which plotted data should be assessed. When an axis only just contains a pair of values, they will generally be considered to be highly divergent. When an axis easily contains these values, they will generally be considered similar, because the difference between values will be dwarfed by the vastness of the scale. Arbitrary rules will not absolve a chart designer's responsibility to consider what their visualization implies (Correll *et al.* 2020).

As Yang *et al.* (2021) discuss, one explanation for these effects draws on Grice's co-operative principle (Grice 1975). This theory, originally concerning linguistic utterances, would suggest that components of a chart, such as axes, will be considered to communicate relevant information about plotted data. Thus, a viewer will derive a designer's intended message from the features of the visualization. Changing one's interpretation of magnitude in accordance with changes to axis range could therefore be considered a coherent response.

Research on risk communication has also explored how visualization design choices affect interpretations of presented information. A set of experiments relevant to the present investigation originated with empirical data which suggested that icon arrays were more effective than text at promoting risk-averse behavior (Stone *et al.* 1997). Further research (Stone *et al.* 2003) suggested that this occurred because the data visualizations only displayed the number of people affected by the negative outcome. Therefore, unlike the text, the icon arrays made the numerator more salient than the denominator (the total number of people in the sample). This was demonstrated empirically in the same study, using bar charts: the difference between numerators (15 vs. 30) appeared much bigger when the larger numerator (30) was used for the upper axis limits, compared to when the denominator (5000) was used for the upper axis limits. Risk reduction (the degree of difference between plotted values) was perceived as smaller when bar charts were extended to incorporate the denominator. Unlike the above studies on y-axis truncation (Pandey *et al.* 2015, Witt 2019, Correll *et al.* 2020, Yang *et al.* 2021, Driessen *et al.* 2022), the lower axis limit was not manipulated, and remained fixed at zero. This pattern of results has been replicated using icon arrays (Garcia-Retamero and Galesic 2010) and pie charts (Hu *et al.* 2014), and a similar effect has been reported for line charts (Taylor and Anderson 1986) suggesting this phenomenon is driven by a common mechanism independent of chart type.

Stone et al.'s (Stone *et al.* 2003) experiment demonstrated that extending the upper limit caused participants to interpret the difference between values as smaller. Unfortunately, the design of this experiment leaves uncertainty as to whether this extension affected interpretations of the magnitude of *the values themselves*, because participants only compared risks between charts in the same condition, not across conditions. However, this issue was addressed by Okan *et al.* (2020), who found that icon arrays which *did not* display the denominator increased perceived risk relative to those which did (with larger increases at smaller probabilities). Including the denominator also resulted in more accurate estimates of the underlying risk probabilities. This accords with the finding that the apparent magnitude of risk decreases when the upper limit is extended in a risk ladder visualization (Sandman *et al.* 1994). This implies that interpretations of magnitude are informed, in part, by the data point's position within the risk ladder's limits.

Encoding Values Using Color

In data visualizations employing geometric encodings (e.g., position, extent), axes are the dimensions along which data are plotted. In colormap visualizations, a different type of axis is present, which is not used to display data directly, but presents the mapping between colors and numerical values, henceforth referred to as a 'color legend'. Default settings in popular visualization tools, such as ggplot2 (Wickham 2016) and Matplotlib (Hunter 2007) tend to employ color legends which use the minimum and maximum values in the data at their extremes. Thus, the potential for values smaller than the minimum, or larger than the maximum, is not encoded by these color legends. This facilitates comparison between values, since using a wide range of colors improves discrimination ability. Crucially, however, it does not facilitate magnitude judgments. Consider, for example, a heatmap showing profits for each quarter over the course of five years. Using the darkest color on the color legend to represent the highest profits could conceal the fact that profits in general have been poor for the entirety of this period, because the color legend is agnostic towards real-world magnitude.

Research involving color legends has often focused on assessing the appropriateness of different color scales and capturing color discriminability through color difference models. Harrower and Brewer (2003) developed a tool for selecting suitable color scales for particular forms of data: sequential scales for ordinal or numerical data, qualitative scales for categorical data, and diverging scales for highlighting midpoints. Other work has identified specific features which make for an effective color scheme, from low-level properties such as uniform luminance (Dasgupta *et al.* 2020) to high-level properties such as consistency with semantic color associations (Lin *et al.* 2013). Researchers have also modeled the impact of mark size on color discriminability (Stone *et al.* 2014) and demonstrated adaptation of color difference models to specific viewing conditions (Szafir *et al.* 2014).

Choropleth maps are one of several types of colormap visualization which map color to numerical data (see also, heatmaps and neuroimaging visualizations). Schiewe (Schiewe 2019) illustrates that impressions of quantity are positively associated with the proportion of a choropleth map occupied by darker colors. The size of geographical regions and the binning of values can both influence the extent to which a map displays colors on the darker end of the chosen color scale, which impacts judgements of presented data. Whilst this study manipulated the appearance of plotted data in maps, other research has held the appearance of plotted data constant in order to study how the context surrounding a color legend affects viewers' inferences. Schloss *et al.* (2019) observed that viewers' spontaneous interpretations of the relationship between color and quantity can depend on which background color is used. Their experiment attempted to reconcile contrasting theories about which aspects of a color stimulus are associated with greater quantities ('dark-is-more'; 'contrast-is-more'; 'opaque-is-more'). They found that viewers associate darker colors with greater quantities when there is no apparent variation in the color scale's opacity. However, when the color scale does appear to have varying degrees of opacity, an 'opaque-is-more' association prevails. For example, black-white color scales appear to have high opacity against a black background (so lighter grays are more readily associated with larger quantities), but low opacity against a blue background (so lighter grays are more readily associated with smaller quantities).

Different interpretations of the same dataset can also arise through modified displays of the same color scale. Empirical research has compared color legends which only use color features (e.g., increasing luminance and decreasing saturation) to indicate uncertainty, to color legends which also signal uncertainty through increasing reduction in the range of possible colors, termed Value-Suppressing Uncertainty Palettes (VSUPs, (Correll *et al.* 2018)). In Correll et al.'s study, participants played a 'Battleship' style game which involved reducing risk by balancing danger and uncertainty. Participants were more likely to favor riskier but more certain options over uncertain options when using VSUPs. Constraining the range of colors at higher uncertainty levels may have reduced the impression that these data points could represent desirable low-danger magnitudes. The experiment we report below examines directly how the range of values in a color legend affects interpretations of magnitude.

Methodology

Outline

The present experiment investigates the influence of color legend range on the cognitive processing of magnitude. We manipulated the color legend's upper bound, such that it was equal to the maximum plotted value (*truncated range*) or it was equal to double the maximum plotted value (*extended range*). We employ the term 'truncated' in a broad sense, referring to a scale that is constrained such that potentially relevant values are omitted, not simply a scale that excludes a zero value. Using a lower bound of zero reduced the number of differences between the two conditions, so that only the upper bound was manipulated. This also meant that plotted values' variability appeared smaller, assisting participants in judging the overall magnitude of these values. For each item, the color palette, geographic regions, and the mapping between colors and numerical values, were identical across conditions. Therefore, the only difference between versions of a given item was the arrangement of the color legend: the map itself remained unchanged.

Rather than asking participants to make abstract judgments about the size of abstract values, we presented fictitious pollution data, and asked how urgently action should be taken to address the pollution levels displayed in each data visualization. This captures participants' assessments of magnitude through the type of judgments which can drive behavior. In addition to increased ecological validity, we also anticipated that pollution data might be able to generate a balanced set of responses to the question of urgency. A variable evoking an extreme negative reaction may have elicited responses at ceiling and one too trivial may have elicited responses at floor. We expected participants to recognize that a sufficient degree of pollution would require action, but also understand that low levels may require less urgent action. We did not provide a specific definition of urgency for participants to use when making their responses. Therefore, different participants' responses may reflect different notions of urgency. However, the within-participants design accounts for individual variation. Each participant's ratings are compared against their own ratings for the alternative condition, allowing for meaningful comparison between conditions.

Pollution levels were displayed in choropleth maps, which use color encoding to display data aggregated at the level of geographic areas. Note that we do not consider the designs of choropleth maps in this experiment to reflect best practice for plotting pollution statistics. Rather, their designs were motivated by the desire to examine the role of color legends in the interpretation of magnitude. Previous research has illustrated that the size of geographical regions can influence ensemble coding in choropleth maps (Schiewe 2019). However, we did not control for this aspect, instead we prioritized ecological validity by using maps with real geographical regions. These maps appeared identical across conditions in order to avoid this bias confounding results.

To control for the possibility that participants used the color legend's numerical labels, rather than the range of values displayed, as a reference for their magnitude judgments, we omitted the color legend's numerical labels in half of trials. This allowed us to test whether the presence of numerical labels affected the degree to which magnitude judgments were influenced by the color legend's upper bound.

Pre-Registration

We predicted that urgency ratings would be higher for truncated legends, compared to extended legends. In addition, we planned to compare whether any difference between these two conditions was moderated by the presence or absence of numerical labels, but made no predictions about existence or direction of any main effect or interaction. Participants completed Garcia-Retamero et al.'s (Garcia-Retamero *et al.* 2016) Subjective Graph Literacy scale, therefore we also planned to test whether any observed effects (or lack of) could be explained by differences in data visualization literacy. This five-item scale is a quick, reliable measure that is correlated with scores on Galesic and Garcia-Retamero's (Galesic and Garcia-Retamero 2011) test-based measure of data visualization literacy. The pre-registration, plus materials, data and analysis code are available at https://osf.io/qe9hf/?view_only=32c420d6ef6c45b1ae2d3dc42dc6fe69.

Design

In each trial, we independently manipulated two aspects of the choropleth map. When the color legend had a *truncated range*, its upper bound was equal to the maximum value displayed in the map. When the color legend had an *extended range*, its upper bound was equal to double the maximum value (and the maximum value displayed in the map appeared at the legend's halfway point). Numerical labels on the color legend were either *present* or *absent*. This resulted in four unique combinations of conditions. We employed a Latin-squared design, ensuring that each participant was exposed to each combination of conditions throughout the experiment, but only saw one combination for each given map. There were a total of 54 trials (48 experimental trials, six attention check trials). Example stimuli are shown in Figure @ref(fig:example-stimuli).

References

Correll, M., Bertini, E., and Franconeri, S., 2020. Truncating the Y-Axis: Threat or Menace? *In*: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, 1–12.

Correll, M., Moritz, D., and Heer, J., 2018. Value-Suppressing Uncertainty Palettes. *In*: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, 1–11.

Dasgupta, A., Poco, J., Rogowitz, B., Han, K., Bertini, E., and Silva, C.T., 2020. The Effect of Color Scales on Climate Scientists' Objective and Subjective Performance in Spatial Data Analysis Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 26 (3), 1577–1591.

Driessen, J.E.P., Vos, D.A.C., Smeets, I., and Albers, C.J., 2022. Misleading graphs in context: Less misleading than expected. *PLOS ONE*, 17 (6), e0265823.

Fischer, J. and Ali, A., 2021. A Federal Ban on Abortion is Wildly Unpopular in All 50 States. *Data For Progress*.

Galesic, M. and Garcia-Retamero, R., 2011. Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making*, 31 (3), 444–457.

Garcia-Retamero, R., Cokely, E.T., Ghazal, S., and Joeris, A., 2016. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making*, 36 (7), 854–867.

Garcia-Retamero, R. and Galesic, M., 2010. Who proficts from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine*, 70 (7), 1019–1025.

Grice, P., 1975. Logic and Conversation. *In*: P. Cole and J.L. Morgan, eds. *Syntax and Semantics Vol.3: Speech Acts*. New York: Academic Press, 41–58.

Harrower, M. and Brewer, C.A., 2003. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40 (1), 27–37.

Hu, T.-Y., Jiang, X.-W., Xie, X., Ma, X.-Q., and Xu, C., 2014. Foreground-background salience effect in traffic risk communication. *Judgment and Decision Making*, 9 (1), 8.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9 (3), 90–95.

Lin, S., Fortuna, J., Kulkarni, C., Stone, M., and Heer, J., 2013. Selecting Semantically-Resonant Colors for Data Visualization. *Computer Graphics Forum*, 32 (3pt4), 401–410.

Okan, Y., Stone, E.R., Parillo, J., Bruine de Bruin, W., and Parker, A.M., 2020. Probability Size Matters: The Effect of Foreground-Only versus Foreground+Background Graphs on Risk Aversion Diminishes with Larger Probabilities. *Risk Analysis*, 40 (4), 771–788.

Pandey, A.V., Rall, K., Satterthwaite, M.L., Nov, O., and Bertini, E., 2015. How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion Techniques. *In*: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. Seoul, Republic of Korea: ACM Press, 1469–1478.

Sandman, P.M., Weinstein, N.D., and Miller, P., 1994. High Risk or Low: How Location on a "Risk Ladder" Affects Perceived Risk. *Risk Analysis*, 14 (1), 35–45.

Schiewe, J., 2019. Empirical Studies on the Visual Perception of Spatial Patterns in Choropleth

Maps. *KN - Journal of Cartography and Geographic Information*, 69 (3), 217–228.

Schloss, K.B., Gramazio, C.C., Silverman, A.T., Parker, M.L., and Wang, A.S., 2019. Mapping Color to Meaning in Colormap Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25 (1), 810–819.

Stone, E.R., Sieck, W.R., Bull, B.E., Frank Yates, J., Parks, S.C., and Rush, C.J., 2003. Foreground:background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes*, 90 (1), 19–36.

Stone, E.R., Yates, J.F., and Parker, A.M., 1997. Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, 3 (4), 243–256.

Stone, M., Szafir, D.A., and Setlur, V., 2014. An Engineering Model for Color Difference as a Function of Size. Boston, Massachusetts: Society for Imaging Science; Technology, 6.

Szafir, D.A., Stone, M., and Gleicher, M., 2014. Adapting Color Difference for Design. Boston, Massachusetts: Society for Imaging Science; Technology, 6.

Taylor, B.G. and Anderson, L.K., 1986. Misleading Graphs: Guidelines for the Accountant. *Journal of Accountancy*, 162 (4), 126–135.

Wickham, H., 2016. *ggplot2*. New York, NY: Springer Science+Business Media, LLC.

Witt, J.K., 2019. Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis. *Meta-Psychology*, 3.

Yang, B.W., Vargas Restrepo, C., Stanley, M.L., and Marsh, E.J., 2021. Truncating bar graphs persistently misleads viewers. *Journal of Applied Research in Memory and Cognition*, 10 (2), 298–311.