# Testing Perceptual Accuracy in a U.S. General Population Survey Using Stacked Bar Charts

Kiegan Rice[1,*], Heike Hofmann[2], Nola du Toit[1], and Edward Mulrow[1]

[1]*NORC at the University of Chicago, Chicago, IL, USA*
[2]*Department of Statistics, Iowa State University, Ames, IA, USA*

## Abstract

The use of visuals is a key component in scientific communication. Decisions about the design of a data visualization should be informed by what design elements best support the audience's ability to perceive and understand the components of the data visualization. We build on the foundations of Cleveland and McGill's work in graphical perception, employing a large, nationally-representative, probability-based panel of survey respondents to test perception in stacked bar charts. Our findings provide actionable guidance for data visualization practitioners to employ in their work.

**Keywords** *data visualization; graphical perception; survey data*

## 1 Introduction

What do viewers see when we show them a data chart? A data chart – at its core – maps values to graphical elements: quantitative elements are represented as measurable features, such as position, size, or shade. Modern data visualizations are much more than a simple, objective mapping of values to a plane; they contain contextual and design elements, and are often structured to support the viewer in understanding a particular view of a set of data or specific pattern underlying the values. Structural choices, such as the choice of a graphical element (e.g., bar or pie), orientation of the chart, and sizing and relative positioning of elements within the chart, are determined by the data visualization practitioner. The design of a data visualization impacts a viewer's ability to achieve the intended understanding; a poorly designed data visualization may leave viewers struggling to understand the content or context, or make it difficult to complete accurate and useful comparisons of values across groups or time points. More broadly, the design of a data visualization can change how viewers interact with the chart.

The comparisons of components is a crucial step in the process of interacting with and understanding the chart. Cleveland and McGill (1984) observed as such, and in their seminal study defined the better visual among a pair as the one that allows viewers to make more accurate comparisons. Based on mappings of quantitative variables to different graphical elements, Cleveland and McGill's study resulted in a ranking of perceptual tasks (such as identifying the larger of two lengths or areas) from most accurate to least accurate, which was then extended by Mackinlay (1986) to a theoretical framework ranking tasks' order along their ordinal and nominal scales.

Cleveland and McGill's work – while a foundational user study in graphical perception – utilized a small convenience sample, consisting of only a few individuals recruited from among

---

the authors' coworkers and their spouses. Heer and Bostock (2010) reproduced Cleveland and McGill's rankings using a larger sample from a crowd sourcing platform, employing a total of 82 Amazon Mechanical Turk (mTurk) workers for the study. Crowd sourced samples were shown by Borgo et al. (2017) to be biased towards more male, younger, and relatively well-educated individuals compared to the general adult U.S. population. Presumably, individuals with a higher education are more likely to have prior exposure to data visualization and might be more familiar with quantitative statements than less well-educated groups within the population; other pre-existing differences between population groups may also drive differences in understanding or interpreting charts. This raises the question of whether conclusions drawn from these studies hold for the general population – a common target audience for data visualization and scientific communication work. Results of polls, surveys, and scientific work are presented visually in the news, shared on social media, posted online or in print materials. It is imperative that data visualizations used to communicate quantitative information are designed with the goal of reaching the entire population. To that end, an inclusive, representative approach to studying how viewers perceive information presented in charts is critical to understand how best to communicate with the general public.

In this work, we seek to answer whether it is possible to reproduce some of the previous findings in the context of a survey with a large, nationally-representative set of respondents. Specifically, we present viewers with structural variations of bar charts and ask them to answer questions comparing the size of elements within those charts. We employ a probability-based survey panel and run a series of perception tests with nationally-representative samples of respondents from that panel. The advantage of using a probability-based approach is two-fold: First, the relatively larger sample of survey participants provides greater statistical power in drawing inference about graphical perceptional abilities. Second, the sample is representative of the general adult public in the U.S., permitting an assessment inhowfar prior results from convenience samples hold with a nationally representative sample and whether there are differences in those results across demographic subgroups. Within that context we focus on the following research questions:

1. How do structural design choices in a data visualization impact viewers' ability to identify the larger of two elements?
2. How is viewer interaction with the task impacted by structural design choices in a data visualization?
3. Are there differences in perception and interaction with the tasks across demographic groups?

To address these questions, we present a series of completed tests and the resulting findings. The remainder of the paper is organized as follows: first, we describe the design of the visual stimuli used in the perception tests. We then describe the population of study respondents and obtained survey sample. Subsequently, we share a summary of the resulting survey responses across each of the tests, including analyses on accuracy of responses and response behavior. Finally, we discuss implications of this work and next steps.

## 2 Study Design

### 2.1 Stimulus

Our use of a survey format guides the format and design of the questions asked and how they are presented to respondents. First, participant instructions must be delivered in a very short and

easily understandable format, because participants cannot ask clarifying questions about the task as they might be able to in a cognitive lab setting. Our tasks are presented as a short series of questions within a larger survey on a variety of topics. Second, we want to utilize content within the stimulus which is not socially or politically charged for the average U.S. adult; this risks participants reacting to the subject matter within the chart rather than focusing on the task. For this reason, we utilize data on living arrangements of older U.S. adults – a topic which most U.S. adults will have some familiarity with, but is not inherently polarizing. Finally, to prevent viewers from being exposed to slight variations of the same stimulus in a row (and risk unforeseen order effects or respondents using prior questions to inform their responses), we either split a survey sample in two and show each subsample a distinct structural version of the chart or test variations of a chart across distinct rounds of the survey. Survey rounds are fielded once a month, so distinct rounds are separated by approximately one month.

Each task is made up of two elements: a visual stimulus and a question about the stimulus. In our study, each visual stimulus is an image of a data visualization, while each question prompts viewers to identify which of two marked pieces in the data visualization is larger.

The comparison between sets of marked pairs is intentionally designed to be a difficult task, with the difference between the values represented in the two marked pieces being chosen close to their just-noticeable difference. The **Just-Noticeable Difference** (JND) is defined as the smallest difference that will be detected 50% of the time. Prior results from studies on bar charts and pie charts (Lu et al. 2022) inform the differences in charts shown to survey panelists.

We employ comparisons at the JND in our tasks in order to maximize our ability to identify the impact of design changes on viewer accuracy and behavior. Asking perception tasks in a survey differs from the controlled environment of a cognitive lab, where these kind of questions may usually be assessed. Rather than asking the same (or similar) type of question with varied signal strength dozens or hundreds of times, we are limited to only a few questions at a time. With a small set of tasks, we need to present tasks that are perceptually hard, and thus ask questions about stimuli that are close to our perceptual threshold. Therefore, we focus on questions which vary the structure of the presented image, but ask viewers to compare the same underlying data values across those varied images.

We ask participants to determine which of two just-noticeably different marked pieces (tiles) is larger within each chart, and focus on three main sets of structural variation in the design of that chart. First, we vary the alignment of the pieces in question. Viewers are presented with two marked pieces in a chart that do not share a common baseline (unaligned pieces), then two pieces that do share a common baseline (aligned pieces). We present the unaligned comparison first to reduce the likelihood that participants recognize that the tasks are identical excepting the alignment and carry over an answer from the aligned task to the unaligned task. Second, we vary the orientation of the chart – we rotate the vertical stacked bar chart, and present a horizontally oriented version of the same chart, with identically sized marked pieces. Finally, we change the aspect ratio of the chart and present a wider version of the horizontally oriented chart which has longer, but thinner, marked pieces.

### 2.1.1   Task 1: Vertical Stacked Bar

Figure 1(a,d) shows the two stacked bar charts shown to participants in the first task. The marked tiles in each plot are 155 pixels apart, which leads to a JND of 3.5 pixels based on Lu et al. (2022)'s model. The heights of the bars are 205 (left) and 213 pixels (right), corresponding to about twice the JND. This difference should lead to a relatively high accuracy rate for
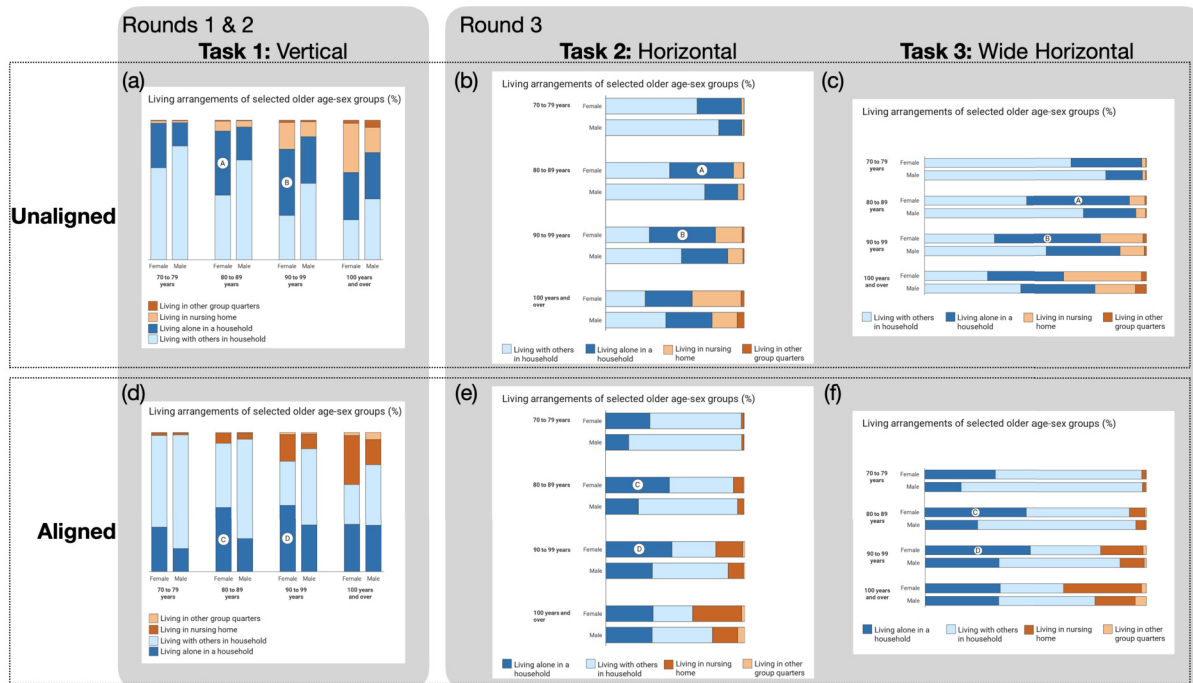
Figure 1: Overview of all of the visual stimuli presented to viewers. In each bar chart, two pieces are marked. The larger piece in each chart are pieces B and D. Please refer to the appendix for a full size image of stimulus (a).

participants and simultaneously limit the amount of frustration resulting from a task that is perceived as 'too hard'.

Both charts show the same data with slight modifications to the order of the levels – the first and second level in each of the bars are reversed between the chart in (a) and the chart in (d). Both charts are displayed at the same size, i.e. in both cases both the difference in size between the bars and the horizontal distance between the bars is the exact same amount. This leaves the vertical positioning of the bars as the only difference between the charts. Any differences in observed responses can therefore be attributed to this difference in presentation.

### 2.1.2   Task 2: Horizontal Stacked Bar

The visual mappings in the second task, shown in Figure 1(b,e), are identical to the first task, but the axes of the chart are rotated so that the stacked bars are represented in a horizontal format. This represents a structural change in how the data are presented to the viewer while preserving the pixel size of the elements viewers must compare. Differences in participant accuracy across Tasks 1 and 2 can thus be assumed to be due to the orientation or layout of the chart rather than the comparison task itself. Tasks 2 and 3 were asked within the same survey round (Round 3) and the full sample was randomly split among respondents, with 50% of participants seeing the Task 2 stimuli and 50% of participants seeing the Task 3 stimuli.

### 2.1.3   Task 3: Horizontal Wide Stacked Bar

The images utilized in the third task, Figure 1(c,f), again represent the same data as in the first two tasks, and the overall image has dimensions which are identical to the images in the first
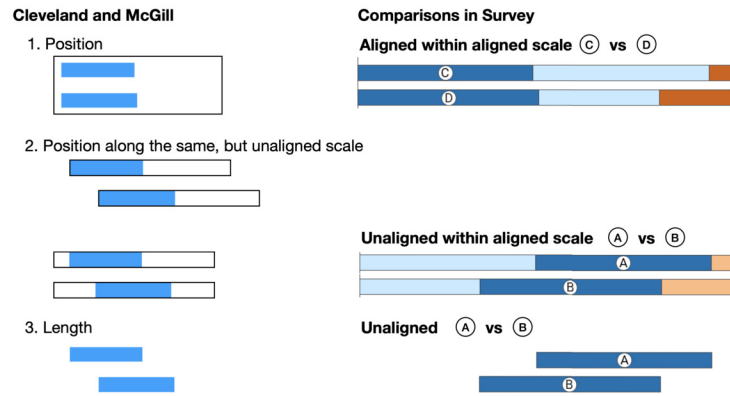
Figure 2: Comparisons made in charts within the Cleveland and McGill ranking (left) and their corresponding representations in our study tasks (right).

two tasks, displaying at the same size during the survey. However, the aspect ratio is adjusted to make the plotting area match that of Task 1, with the mapping area being wider rather than taller. The length of bars is increased to fit the new dimensions. This increases the difference in the length of bars to 13 pixels (previously 8 pixels). This increase in width allows us to test whether participant accuracy is higher when the elements in the comparison task have a larger visual difference in length relative to Tasks 1 and 2. The widths of the tiles are adjusted accordingly (from 50 to 30 pixels) to keep the overall area of the tiles approximately constant.

### 2.1.4   Question Text

When viewing each chart, participants were asked to compare the relative sizes of marked elements within the chart:

"There are many charts used in the news media to portray data visually. Looking at the chart below, which of the marked dark blue pieces is bigger, A or B? Just your best guess is fine. [] A is bigger; [] B is bigger; [] They are the same."

When presented with the aligned version of each chart, pieces were marked with a C and D and the question text is updated accordingly. In all scenarios, the second option (B [D] is bigger) is the correct response.

For a given task, viewers are first presented with the unaligned version of the task, followed by the aligned version of the task. The time in seconds that each respondent spent on each task was recorded, as well as whether the participant zoomed in on each chart while answering the question. Respondents were also asked to rate their certainty in their response to each question on a five-point scale.

### 2.1.5   Alignment

What we call 'aligned' and 'unaligned', here, is similar to Cleveland and McGill's positions along aligned and unaligned scales, but with some modifications. Figure 2 gives an overview of the comparisons of tasks 1 through 3 and the closest corresponding tasks in Cleveland and McGill. In this paper, both 'aligned' and 'unaligned' bars share the same axis.

Aligned tiles are additionally anchored in the same position along the axis, i.e. the difference between their sizes can be reduced to a positional assessment, which in the absence of three-dimensional cues seems to be the general strategy viewers use when evaluating bar charts (Zacks

et al. 1998). Unaligned tiles do not share this anchor of position, i.e. the difference between their sizes needs at least two comparisons of positions. However, unaligned bars in a stacked bar chart are not just assessments of the lengths of bars – the other tiles in the chart provide a reference frame of the shared axis, which *should* help with an assessment of the tiles' sizes. As shown in the sketch of Figure 2, this framing leads to a stimulus that is between Cleveland and McGill's task of 'position along the same but unaligned axis' and an assessment of length.

We would expect that comparing unaligned tiles is a harder task (with correspondingly lower levels of accuracy) than a comparison of aligned tiles, with the framing given by the other tiles in the same column mitigating some of this difficulty. In particular, the additional context of the other tiles should *help* rather than *hurt* the comparison, unlike other situations, where additional information besides the pieces involved in the comparison were found to be distractors (Talbot, Setlur, and Anand 2014). Participants first saw the chart with the unaligned comparison, followed by the chart with the aligned comparison, but could not return to the first comparison after completing it. While there is some risk of participants recognizing the subsequent tasks to be the same value comparisons, the unaligned (harder) task was shown first to reduce the likelihood of participants carrying over an answer from the aligned (easier) task.

## 2.2 Participants

Participants were recruited as part of NORC's AmeriSpeak panel,[1] which utilizes a probability-based sampling methodology and samples U.S. households from NORC's National Sample Frame that provides coverage of over 97% of U.S. households. The current panel size is 54,001 panel members aged 13 and over residing in over 43,000 households (Dennis 2019, updated 2022). Each test was conducted using the AmeriSpeak Omnibus survey, which runs biweekly and samples around 1,000 U.S. adults to answer questions on a variety of topics. Our tests require visual inspection of an image; for this reason, our survey questions were only presented to web-based panelists and not panelists who respond via phone interviews. The panel has 45,565 web-based panelists, representing 93.3% of the total panel and 96.3% of the total panel weights.

The Omnibus sample is not longitudinal in nature, i.e. we do not have any information about whether the same panelists were included in multiple rounds. While there is a (small) chance that panelists are included in multiple rounds of our data, there was at least a one month gap between viewing each visual stimulus, and for the purposes of the analysis we will consider data from these viewings as independent.

## 2.3 Survey Weighting

The sample used for each Omnibus study is selected from the AmeriSpeak Panel using 48 sampling strata, split by age, race/Hispanic ethnicity, education, and gender, with the size of each sample per sampling stratum determined by the population distribution for each stratum. In addition, sample selection takes into account expected differential survey completion rates by demographic groups in order to achieve a representative sample of the target population. Collected data are weighted to the latest Current Population Survey (CPS) benchmarks from the U.S. Census Bureau. They are balanced by gender, age, education, race/ethnicity, and geographic region. All calculations in this paper are done in R (R Core Team 2022), and weights are applied

---

[1]NORC is a non-partisan organization that collects data for government and non-government clients.

Table 1: Survey rounds: dates, number of panelists (nominal sample size), effective sample size, sum of weights, and factors for the combining of surveys.

| Name | Date | # Panelists | Effective Sample Size | Sum of weights $\sum_i w_i$ | $\lambda_i$ |
|------|------|-------------|----------------------|------------------------------|-------------|
| Round 1 | April 2022 | 933 | 521.1 | 934.9 | $\lambda_1 = 0.343$ |
| Round 2 | May 2022 | 953 | 485.7 | 953.4 | $\lambda_2 = 0.320$ |
| Round 3 | June 2022 | 921 | 513.5 | 923.1 | $\lambda_3 = 0.338$ |
| **Combined** | — | 2807 | 1520.8 | | |

in analyses using the `survey` package (Lumley 2004) version 4.0 (Lumley 2020) based on Lumley (2010).

When combining responses from different surveys, weights are rescaled with respect to the total population, so that their order after combining still reflects their relative importance. We *combine* (rather than cumulate) a set of $\ell$ surveys $S_1, S_2, \ldots S_\ell$ (with $\ell \geqslant 2$) as described in O'Mircheartaigh and Pedlow (2002), by multiplying weights in $S_i$ by $\lambda_i$, for $1 \leqslant i \leqslant \ell$. $\lambda_i \in [0, 1]$ with $\sum_i \lambda_i = 1$ is given as

$$\lambda_i = \frac{n_i/d_i}{\sum_{j=1}^{\ell} n_j/d_j}, \tag{1}$$

where $n_i$ is the nominal sample of survey $S_i$ and $d_i$ are the design effects for the estimators. Here, $d_i$ are estimated as

$$d_i = 1 + CV(w \in S_i)^2, \tag{2}$$

where $CV$ is the coefficient of variation of the weights $w$ within each survey sample $S_i$, and is estimated as in Kish (1965):

$$CV(w \in S) = \frac{\widehat{Var(w)}}{\bar{w}^2}, \tag{3}$$

and where $\bar{w}$ and $\widehat{Var(w)}$ are sample mean and sample variance of the weights $w$ of survey sample $S$, respectively.

The data for this paper were collected in several rounds. The resulting number of panelists, effective sample sizes, and $\lambda$ values are shown in Table 1.

## 3   Results

### 3.1   Respondents

A total of 2807 panelists participated across the three rounds. The number of responses and corresponding effective sample sizes in each round are shown in Table 1. All responses were combined into one set of survey responses, with adjusted combined sample weights and indicators for which task each respondent was exposed to. The resulting sample sizes corresponding to each task are shown in Table 2.

Distributions of demographic characteristics across each task are shown in Figure 3, the black points and error bars show percentages for US adults based on the 5 year estimates of the American Community Survey 2021. The distribution of demographics are all quite similar

Table 2: Survey tasks: number of participants (nominal sample size) and effective sample size for each task.

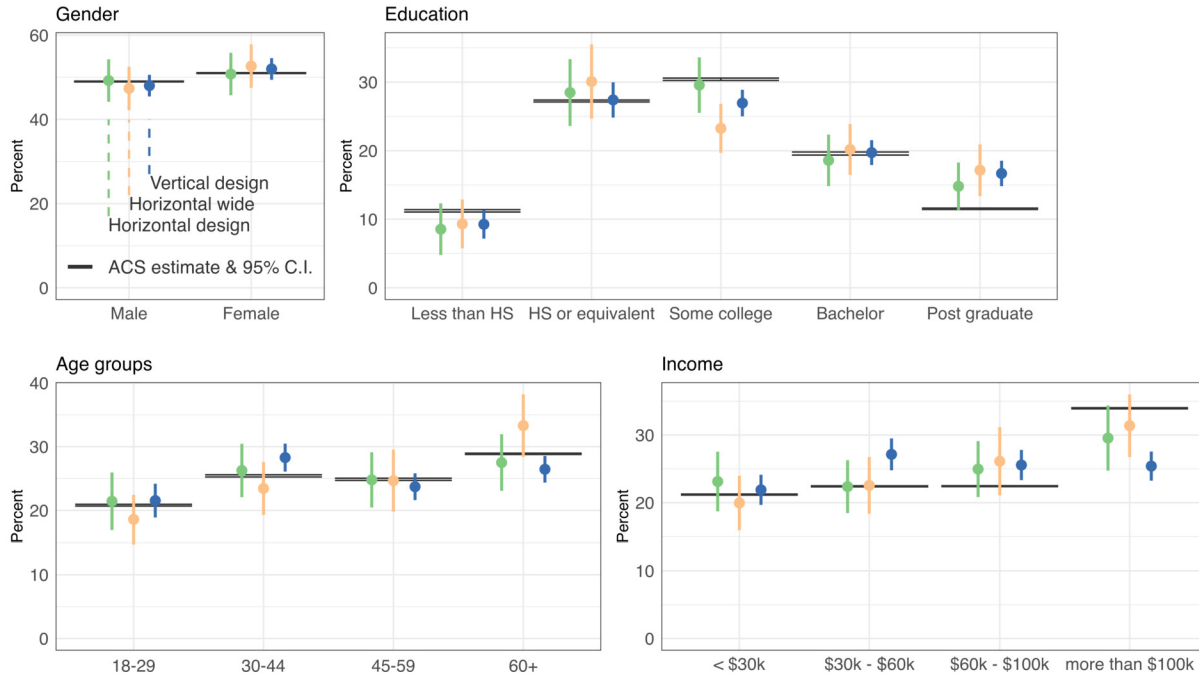| Name | Description | # Participants | Effective Sample Size |
|---|---|---|---|
| Task 1 | Vertical | 1886 | 1007.1 |
| Task 2 | Horizontal | 459 | 266.5 |
| Task 3 | Horizontal wide | 462 | 249.1 |



Figure 3: Overview of participants' demographics in each of the tasks: the colored dots show marginal percentages (with 90% margins of error as vertical lines in the same color) for demographics across the three tasks. In comparison, the dark lines show upper and lower 95% confidence of these demographics as estimated by the American Community Survey (ACS) 2021. Note that for a lot of these intervals the estimates are so close together that they appear as a single line.

across tasks and their 90% margin of errors (vertical lines) for the most part cover the Census estimates.

## 3.2  Accuracy of Responses

We first investigate respondent accuracy in selecting the correct response. Remember, that participants were able to select the option 'They are the same' as response. This leads us to several ways to model accuracy and response. For the purposes of practical interpretation, the argument could be made that 'they are the same' is a correct choice, as the options are visually very similar and not substantially different values within the context of the data shown in the chart. However, we are interested in understanding whether viewers *can* perceive the difference and
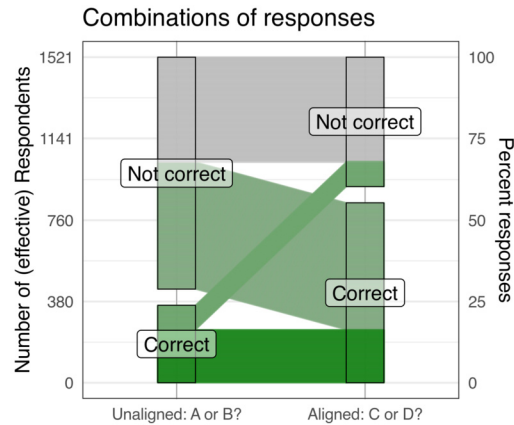
Figure 4: Participants' response pattern to both questions. The coloring reflects the number of correct responses (two in full green, zero in grey, and one correct answer in grey-green). When tiles are aligned along the same axis, more than twice the number of respondents correctly identifies the larger tile. The pattern in responses for participants who only answered one of the questions correctly is asymmetric; that is, a much higher percentage of participants answered the unaligned task incorrectly and the aligned task correctly than answered the unaligned task correctly and the aligned task incorrectly.

correctly identify which piece is larger and thus consider multiple ways of modeling response to investigate participant response patterns. We begin by defining a measure of binary 'correctness', for which all answers that are not the correct option (B [D] is bigger) are 'incorrect', including the selection of 'they are the same'.

Figure 4 displays all responses along the binary correctness measure, separated by whether the stimulus was an aligned task or unaligned task. We can see that levels of accuracy for all responses are significantly higher for the 'easier' (aligned) task, with about twice as many respondents correctly selecting the larger of the two marked elements.

Because each participant was shown both the aligned and unaligned versions of the chart, we can use a paired $t$-test to compare mean accuracy between the two charts. The resulting $t$-statistic is highly significant ($t$ statistic: 21.2, df: 2265, $p$-value: $\ll 0.0001$).

Next, we consider an ordinal model to investigate response behavior across all three options – 'A [C] is bigger', 'B [D] is bigger', and 'They are the same', and consider these response patterns across each of the stimuli.

Figure 5 shows the results of a cell-means model with ordinal response $Y_k$, where $Y_k$ is the $k$th response, $1 \leqslant k \leqslant N$ for $N$ number of (nominal) responses. We use an encoding of $Y_k \in \{1, 2, 3\}$ for 'correct' ($\ell = 1$), 'they are the same' ($\ell = 2$), and 'wrong' ($\ell = 3$):

$$\text{logit } P(Y_k \leqslant \ell) = \mu_{ij\ell(k)}, \tag{4}$$

where $i \in \{1, 2\}$ is the comparison type (1 = Aligned, 2 = Unaligned), and $j \in \{1, 2, 3\}$ is the chart design, with 1 = Vertical, 2 = Horizontal, and 3 = Horizontal wide. Note that we are using the notation $ij\ell(k)$ to emphasize the functional relationship between the $k$th response and the model parameters while avoiding double-subscripts, i.e. $ij\ell(k)$ signifies the index combination of $i, j$, and $\ell$ of the $k$th response, i.e. $ij\ell(k)$ is 123, if the $k$th response was a wrong response ($\ell = 3$) for an aligned comparison ($i = 1$) in a horizontal ($j = 2$) barchart.
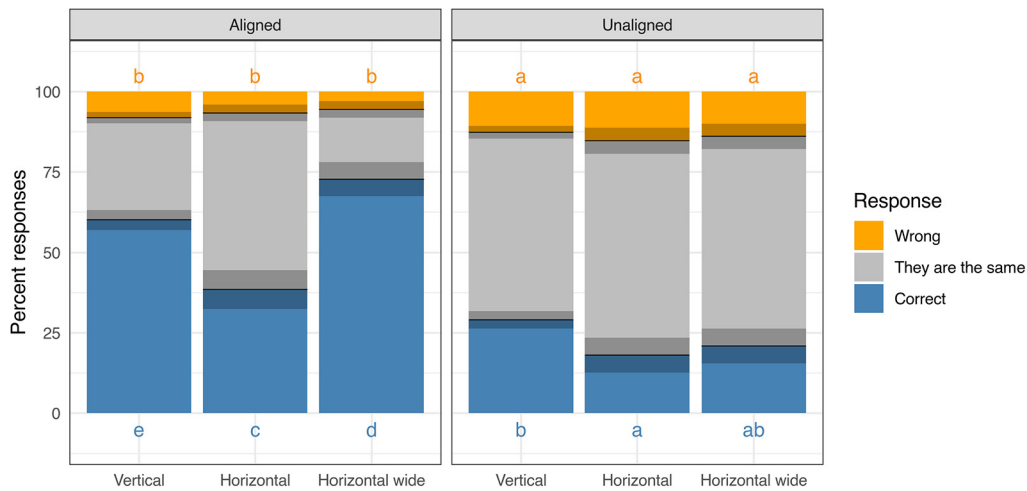
Figure 5: Responses by task across the designs. The overlaid rectangles represent 95% confidence intervals. The letters in blue and orange encode significances between pairwise proportions (Piepho 2004): two bars have a significantly different proportion (at a 5% significance level) if they do not have a letter in common. There is no significant difference between the designs for wrong responses.

The letters in the figures and tables allow us to assess whether differences between two means are significant: in the tables, two estimates within the same column are significantly different (at the 5% level) if they do not have any letter in common. The same is true for letters along the top and bottom of the figures. This method is called connected letters and has been introduced by Piepho (2004). In Figure 5, we see, that the proportions of wrong responses are significantly different between tasks, but designs within each task are not significantly different. For correct responses, all proportions are significantly different except for the proportion of correctly assessing the difference between unaligned rectangles in the wide horizontal design: their label of `ab` shows us that assessing the differences between unaligned rectangles in the wide horizontal design is not significantly better than in the horizontal design (labelled `a`), but also not significantly worse than in the vertical design (labelled `b`). The estimated values and 95% confidence intervals are shown in Table 3.

One pattern in accuracy holds across each of the three structural variations: **the aligned task has a higher level of accuracy than its unaligned counterpart**. Interestingly, while we expect an improvement in accuracy when shifting from the horizontal to the horizontal wide design given the larger difference in pixel length between the two pieces, the resulting effects on the accuracy of the responses are not completely straightforward: the shift from a vertical to the (tall) horizontal design is detrimental to an accurate perception for both aligned and unaligned comparisons. The re-scaled design of the wide horizontal bars reclaims some of the loss for unaligned bars and outperforms the vertical design by a similar margin in aligned bars, but does not out-perform the vertical design when comparing unaligned tiles.

Rates of selecting the response 'wrong' are low across all three structural variations and the aligned and unaligned tasks; while viewers select the response 'wrong' at significantly higher rates for all three unaligned tasks relative to the aligned tasks, those rates do not differ significantly across the three structural variations. Most of the observed differences in response patterns across structural variations are attributed to respondents' selection between the correct option

Table 3: Estimated odds from the cell-means model for response patterns. Letters behind numbers indicate pairwise significances. Within the same column, values are significantly different (at a 5% level of significance) if they do not share the same letter.

| Odds of selecting responses by task and chart type | | | | | | |
|---|---|---|---|---|---|---|
| | correct vs. same or wrong | | | correct or same vs. wrong | | |
| | Est. | [95% CI] | | Est. | [95% CI] | |
| **Unaligned** | | | | | | |
| Horizontal | 0.22 | [0.15, 0.32] | a | 5.54 | [4.04, 7.59] | a |
| Horizontal wide | 0.26 | [0.19, 0.37] | ab | 6.20 | [4.46, 8.61] | a |
| Vertical | 0.41 | [0.36, 0.47] | b | 6.90 | [5.75, 8.28] | a |
| **Aligned** | | | | | | |
| Horizontal | 0.63 | [0.49, 0.81] | c | 14.02 | [9.23, 21.30] | b |
| Horizontal wide | 2.67 | [2.04, 3.48] | d | 17.12 | [10.49, 27.95] | b |
| Vertical | 1.51 | [1.33, 1.71] | e | 11.33 | [8.99, 14.28] | b |

or the 'they are the same' option.

To better comprehend these observed patterns, we investigate viewer interaction with the tasks more broadly as well as differences in response selection across demographic groups.

### 3.3   Respondent Behavior

One contributing factor to the observed patterns in response accuracy might be the way that participants interact with the different designs. Across all tasks, about half of all participants make use of the option to zoom into charts. We observe that while zooming does help with the overall accuracy (which is in agreement with the findings by Lu et al. (2022) about the physical size of stimuli), the increase is not significant. However, different designs lead to different rates of zooming: we observe in Figure 6 that a significantly lower rate of respondents chose to make use of zooming when answering the horizontal designs compared to the vertical design.

To formalize this pattern in a model, let $Y_k$ describe the zooming behavior on the $k$th response. We model zooming behavior (no = 0, yes = 1) as a logistic regression by correctness of response ($\rho$), task ($\tau$), and design ($\delta$) of the chart:

$$\text{logit } P(Y_k \leqslant 1) = \mu + \rho_{\ell(k)} + \tau_{i(k)} + \delta_{j(k)}. \tag{5}$$

The resulting model estimates and confidence intervals are displayed in Table 4. We observe that zooming behavior *only* differs significantly by structural design; rates do not differ significantly between the aligned and unaligned tasks, nor do they differ significantly by the chosen response. This higher rate of zooming does not necessarily lead to higher accuracy; although rates of accuracy are significantly higher for the vertical orientation relative to the horizontal orientation, they are not significantly higher – and in fact, are significantly lower on the aligned task – than the accuracy for the horizontal wide orientation. Using linear scores for the response of certainty, with 'not certain at all' assigned a score of 1 and 'extremely certain' assigned a score of 5, we can estimate the effects of task, chart design, and correctness on certainty by using
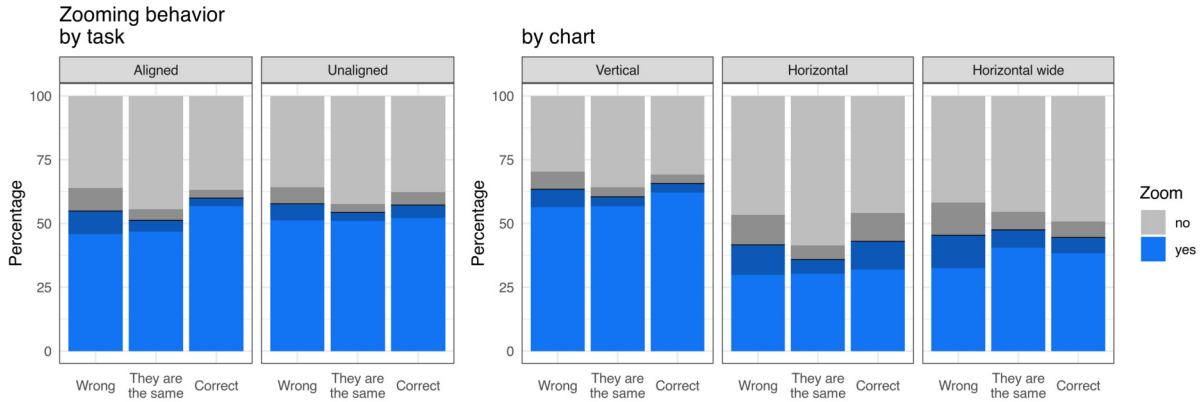
Figure 6: Rates of zooming by task, chart design, and correctness. The overlaid rectangles represent 95% confidence intervals. There are significantly higher rates of zooming by respondents for the vertical task compared to both horizontal tasks; rates of zooming do not differ by correctness of response or alignment in the stimulus.

Table 4: Coefficients for logistic regression of zooming by task.

Estimates for logistic regression on zooming behavior

| term | estimate | SE | t-statistic | p-value |
|---|---|---|---|---|
| $\widehat{\mu}$ | $-0.35$ | 0.15 | $-2.4$ | 0.0155 |
| Response | | | | |
| They are the same $\widehat{\rho}_2$ | $-0.17$ | 0.09 | $-1.9$ | 0.0578 |
| Wrong $\widehat{\rho}_3$ | $-0.06$ | 0.13 | $-0.5$ | 0.6466 |
| Task | | | | |
| Aligned $\widehat{\tau}_2$ | 0.00 | 0.06 | $-0.1$ | 0.9435 |
| Chart design | | | | |
| Horizontal wide $\widehat{\delta}_2$ | 0.27 | 0.15 | 1.8 | 0.0752 |
| Vertical $\widehat{\delta}_3$ | 0.98 | 0.13 | 7.7 | $< 0.0001$ |

a cell-means model of the form:

$$Y_k = \mu_{ij\ell(k)} + \epsilon_k, \tag{6}$$

where $1 \leqslant k \leqslant N$ for $N$ responses, $\mu_{ij\ell(k)}$ is average certainty (measured on a scale from 1 to 5) of the four combinations of task and correctness by each of the three designs, for task $i = 1, 2$ (unaligned/aligned), design $j = 1, 2, 3$ encoded as horizontal, wide horizontal, and vertical stacked bar charts, and participant's decision $\ell = 1, 2$, encoded as correct or wrong, respectively. We also assume that errors are normally distributed, i.e. $\epsilon_k \overset{i.i.d}{\sim} N(0, \sigma^2)$ for all $1 \leqslant k \leqslant N$. The results are shown in Figure 7. We find that the highest scores for certainty are associated with the aligned task, but certainty scores are not significantly different between correct and wrong responses. Scores are (mostly) significantly lower for the unaligned task.
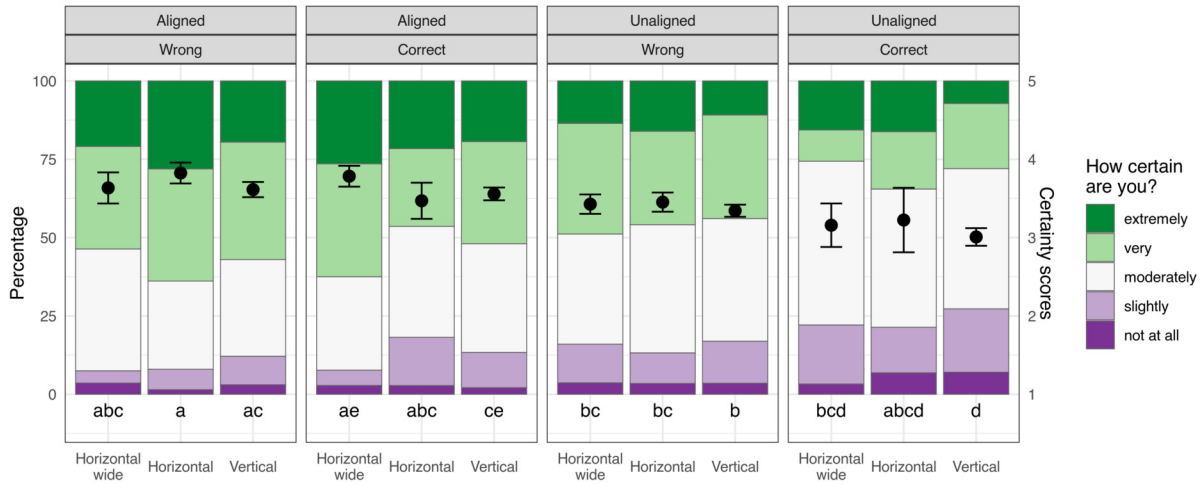
Figure 7: Certainty by task, chart design, and correctness. Correct answers in the unaligned task have the lowest certainty scores associated with them. Incorrect responses to the unaligned task when viewing the vertical stacked bar are associated with a significantly higher certainty score. Error bars around the points show estimated scores (on a scale of 1 to 5) with corresponding 95% CI. The letters underneath the bars indicate significance at 5%. The scores for two bars are significantly different if they do not share a letter.

Interestingly, the lowest certainty scores are associated with correct responses to the unaligned task; although respondents successfully complete the task, they are less certain in their selection.

## 3.4 Differences Across Demographic Groups

Finally, we turn to investigating responses across demographic groups, a core benefit provided by the large and representative nature of our survey samples. Figure 8 displays response patterns across our three structural variations by gender, age (4 groups), educational attainment (5 groups), and income level (4 groups). Our pattern of higher accuracy in selecting the correct response on the aligned task relative to the unaligned task holds across all demographic groups and structural variations. However, we do observe different response patterns, particularly on the unaligned task, across demographic groups.

Let $Y_k$ be the response of participant $k$, on a scale from $1 = $ 'wrong', $2 = $ 'they are the same' to $3 = $ 'correct'. We use a generalized cumulative logistic regression, where $\mu_\ell$ are intercepts $1 \leqslant \ell < 3$, $X_k$ are demographics of the $k$th participant (the $k$th row of the corresponding model matrix), and $\beta_\ell$ is a vector of coefficients:

$$\text{logit } P(Y_k \leqslant \ell \mid X_k) = \mu_\ell + X_k \beta_\ell. \tag{7}$$

Note, that $\beta_\ell$ encompasses all estimable parameters of the demographic variables gender, age category, education category and income category for a total of eleven parameters. The resulting model estimates and confidence intervals are shown in Table 5.

When considering the easier (aligned) task, response patterns do not differ significantly by age, gender, or education level. We do observe, however, a significantly higher log odds of selecting the 'correct' or 'they are the same' response (or significantly lower log odds of selecting 'incorrect') among those with an income between \$60,000 and \$100,000. When we consider the

(a) Comparisons of aligned tiles
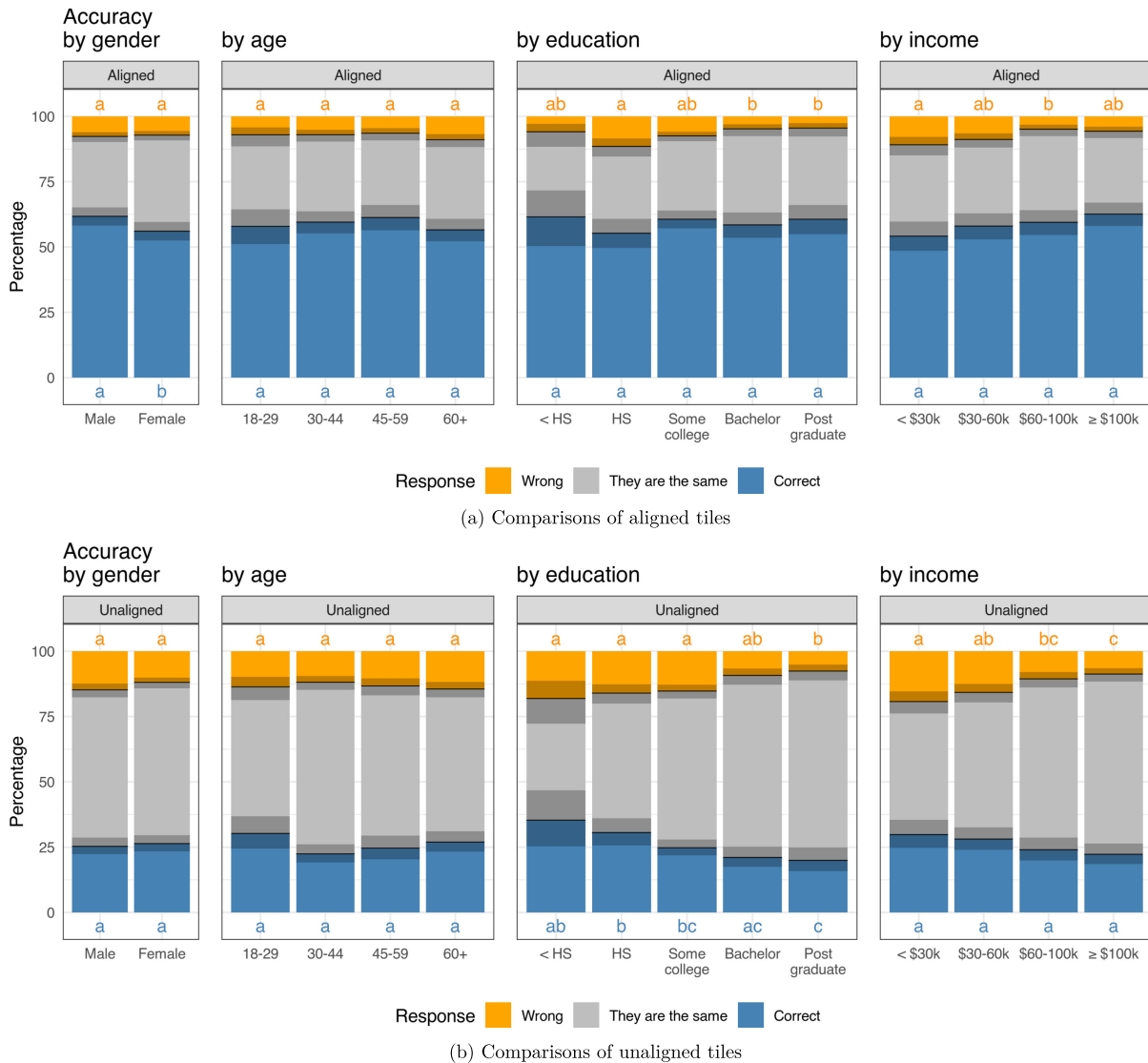


(b) Comparisons of unaligned tiles

Figure 8: Response patterns by alignment and demographic levels. For aligned tiles, age and education are not significant factors. However, gender and income levels do have a (small) effect. For the more difficult task of comparing unaligned tiles, higher levels of education and higher levels of income are associated with a significant increase of panelists choosing a response of 'they are the same', resulting in significant decreases of wrong answers for higher education attainment and higher income groups, and significant decreases in correct answers for higher levels of education.

more difficult (unaligned) task, we see significant separation in response patterns across gender, educational attainment, and income groups. We do not see significant differences in responses by age for either task. Interestingly, we see lower log odds of selecting the 'correct' response among those with higher educational attainment (those with a bachelor's degree or post graduate study) *and* lower log odds of selecting the 'incorrect' response; as income and educational attainment increase, respondents are more likely to select the 'they are the same' option during the difficult

Table 5: Demographics matter for perception, particularly when the tasks get harder.

Odds of accuracy by task and demographics of respondents

| | Aligned tiles | | | | | | Unaligned tiles | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | correct vs. same or wrong | | | correct or same vs. wrong | | | correct vs. same or wrong | | | correct or same vs. wrong | | |
| | Est. | [95% CI] | | Est. | [95% CI] | | Est. | [95% CI] | | Est. | [95% CI] | |
| Intercept | | | | | | | | | | | | |
| | 1.51 | [0.91, 2.50] | | 12.75 | [5.59, 29.10] | *** | 0.60 | [0.35, 1.03] | . | 3.24 | [1.70, 6.19] | *** |
| Gender | | | | | | | | | | | | |
| Female | 0.82 | [0.67, 1.01] | . | 1.19 | [0.81, 1.74] | | 1.03 | [0.82, 1.30] | | 1.42 | [1.07, 1.90] | * |
| Age | | | | | | | | | | | | |
| 30-44 | 1.08 | [0.78, 1.49] | | 0.87 | [0.43, 1.75] | | 0.79 | [0.55, 1.13] | | 0.94 | [0.59, 1.49] | |
| 45-59 | 1.13 | [0.80, 1.60] | | 0.95 | [0.46, 1.95] | | 0.92 | [0.62, 1.35] | | 0.76 | [0.48, 1.22] | |
| 60+ | 0.96 | [0.69, 1.34] | | 0.70 | [0.35, 1.40] | | 1.01 | [0.70, 1.45] | | 0.73 | [0.46, 1.15] | |
| Education | | | | | | | | | | | | |
| HS or equivalent | 0.77 | [0.46, 1.30] | | 0.50 | [0.20, 1.21] | | 0.81 | [0.47, 1.40] | | 1.22 | [0.66, 2.26] | |
| Some college | 0.91 | [0.56, 1.49] | | 0.74 | [0.32, 1.71] | | 0.63 | [0.38, 1.06] | . | 1.17 | [0.65, 2.12] | |
| Bachelor | 0.79 | [0.48, 1.32] | | 1.12 | [0.42, 2.97] | | 0.53 | [0.31, 0.93] | * | 1.82 | [0.91, 3.64] | . |
| Post graduate | 0.84 | [0.49, 1.44] | | 1.16 | [0.41, 3.30] | | 0.51 | [0.28, 0.91] | * | 2.18 | [1.04, 4.57] | * |
| Income | | | | | | | | | | | | |
| $30k - $60k | 1.19 | [0.87, 1.63] | | 1.25 | [0.77, 2.05] | | 1.00 | [0.71, 1.41] | | 1.22 | [0.82, 1.82] | |
| $60k - $100k | 1.23 | [0.90, 1.68] | | 2.14 | [1.18, 3.86] | * | 0.87 | [0.60, 1.27] | | 1.88 | [1.24, 2.87] | ** |
| more than $100k | 1.36 | [0.99, 1.88] | . | 1.57 | [0.86, 2.86] | | 0.86 | [0.59, 1.25] | | 2.18 | [1.34, 3.56] | ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

task. This speaks to the difficulty of the task, and respondents' interpretation of the task when it is more difficult to perceive small differences.

## 4  Conclusions

Through this work, we have tested the use of probability-based survey panels to ask perception questions; within a large survey covering a variety of topics, and with a limited number of questions for our specific tasks, we are able to measure viewer perception and produce results consistent with prior studies. Testing data visualization design structures and viewer behavior on a nationally-representative sample is of broad interest and applicability to the scientific communication community, and this study demonstrates a framework under which to complete further work in this area.

With our survey framework, we have shown that we can rank structural variations on design and tasks by their relative accuracy; we identified significantly different rates of respondent accuracy under different designs presenting the same data. Our study sample is larger than prior studies with a similar scope. That large sample size improves our power to identify this signal.

Further, we can also study respondent behavior and how viewers interact with the chart. Paradata on viewer interaction with the chart (e.g., zooming and time spent on each task) can be collected as well as directly asking respondents questions about their certainty. Not only can we collect this information, but we can also identify significant differences in this behavior across different designs. Understanding viewer interaction and engagement is a key component to designing effective data visualizations. If a particular design leads viewers to zoom in more to investigate it when determining a response, we might consider that design to be more frustrating to viewers 'in the wild' when viewing it outside of the context of our tests.

Perhaps most salient among our results are the patterns in response behavior across demographic groups. The large size of our sample, paired with the probability-based survey panel approach, afford us the ability to dive into response patterns among demographics across tasks and identify significant differences among them. In particular, the differences across educational attainment groups and income groups underscore the importance of utilizing a representative population when testing perception; prior results in this area may be absorbing bias in responses due to the larger representation of individuals with higher education levels among those study respondents. If studies using non-representative or crowd-sourced samples find non-significant results when studying elements of design, those results may not actually be non-significant for other populations, particularly lower education or lower income populations. Researchers should not disregard non-significant findings in those survey studies but rather ask whether the lack of significance would hold across population subgroups.

Our study results provide actionable insights that data visualization practitioners can utilize in the design of stacked bar charts. Comparisons of interest should be lined up and share a common baseline within a chart; in an interactive setting, it may be beneficial to give viewers the possibility of re-ordering categories to support effective comparison across groupings. Allowing users to zoom in on images may not impact perceptual accuracy, but it leads to higher certainty in viewers' comparisons. Considering the target audience for visualizations and what key findings are most important to convey should be an important step in the process of designing a data visualization.

While our results are limited to mapping in a stacked bar chart with a specific topic, they demonstrate among them a wealth of insights about how respondents perceive and interact with

data visualizations. More expansive work on this topic should be completed to understand how the general public interacts with and understands charts, including expanding to a wider set of structural variations, more aesthetic design variations, and considering differences in respondent behavior across different data topic areas.

## Supplementary Material

The source code and data used in this paper and an example stimulus image are available on a GitHub repository at https://github.com/kiegan/testing-charts-jds.

## References

Borgo R, Lee B, Bach B, Fabrikant S, Jianu R, Kerren A, et al. (2017). Crowdsourcing for information visualization: Promises and pitfalls. In: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (D Archambault, H Purchase, T Hoßfeld, eds.), In: *Lecture Notes in Computer Science.*, 96–138. Springer International Publishing.

Cleveland WS, McGill R (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387): 531–554. https://doi.org/10.1080/01621459.1984.10478080

Dennis J Michael (2019), updated 2022. "Technical Overview of the AmeriSpeak Panel NORC's Probability-Based Household Panel." NORC at the, 2019, updated 2022.

Heer J, Bostock M (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In: *Conference on Human Factors in Computing Systems – Proceedings*, volume 1, 203–212. ACM. https://doi.org/10.1145/1753326.1753357

Kish L (1965). *Survey Sampling*. Wiley.

Lu M, Lanir J, Wang C, Yao Y, Zhang W, Deussen O, et al. (2022). Modeling just noticeable differences in charts. *IEEE Transactions on Visualization and Computer Graphics*, 28(1): 718–726. https://doi.org/10.1109/TVCG.2021.3114874

Lumley T (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1): 1–19.

Lumley T (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley and Sons.

Lumley T (2020). "Survey: Analysis of Complex Survey Samples".

Mackinlay J (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2): 110–141. https://doi.org/10.1145/22949.22950

O'Muircheartaigh C, Pedlow S (2002). Combining samples vs. cumulating cases: A comparison of two weighting strategies in NLSY97. In: *ASA Proceedings of the Joint Statistical Meetings*, 2557–2562. http://www.asasrms.org/Proceedings/y2002/Files/JSM2002-001082.pdf.

Piepho H-P (2004). An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2): 456–466. https://doi.org/10.1198/1061860043515

R Core Team (2022). *R: A Language and Environment for Statistical Computing. Manual.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Talbot J, Setlur V, Anand A (2014). Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320.

Zacks J, Levy E, Tversky B, Schiano DJ (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology. Applied*, 4(2): 119–138. https://doi.org/10.1037/1076-898X.4.2.119.