Article Type: Special Issue on Inclusive Data Experiences

Measuring Real-World Understanding of Patterns in Data Graphics

Kiegan Rice, *NORC at the University of Chicago, Chicago, IL, 60603, USA*
Sydney Bell, *NORC at the University of Chicago, Chicago, IL, 60603, USA*
Taylor Wing, *NORC at the University of Chicago, Chicago, IL, 60603, USA*
Heike Hofmann, *Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, 68588, USA*
Nola du Toit, *NORC at the University of Chicago, Chicago, IL, 60603, USA*

Abstract—Presenting data in a visual format, including charts and data graphics, is a cornerstone of effective science communication. While many prior studies have investigated humans' ability to effectively perceive values in charts, fewer have focused on the translation of perceived values to real-world conclusions. Those that do focus on real-world understanding often utilize convenience samples or focus on very simple graphic formats, resulting in an incomplete understanding of how viewers translate data graphics into meaningful conclusions. We utilized a nationally-representative, probability-based sample of over 3,000 participants in the United States across 3 rounds to test user understanding of the content presented in charts. We find that both educational attainment and age play a role in accuracy of context-dependent value estimation in a chart as well as assessing whether presented statements are supported by a chart. We also find that strong ability on perception items increases ability on assessing validity of presented statements, but the converse is not necessarily true. Our study findings demonstrate a need for further study on how chart comprehension and comfort with drawing real-world conclusions differs across demographic groups and across commonly-used chart types. Additionally, this work highlights that complex charts can be inaccessible to viewers who lack confidence in reading and interpreting a chart.

Data visualization displays data in visual elements and maps numbers to shape, position, color, size, etc to help illustrate relationships and patterns in the data. It is a valuable form of scientific communication as viewers use the resulting graphics to make real-life decisions based on data.

We know a lot about perception, or how people perceive data in charts. For example, there have been user studies on which shapes result in the most accurate relative size comparison [1], [2], the smallest visual difference detectable in a chart [3], [4], the impact of alignment of visual elements on viewer perception [5], and how a change in scale impacts perception of pattern magnitude [6], [7]. In concert, these user studies provide rich information about perception of value mappings in graphs. There have been several works which aggregate and summarize findings from perception studies on basic visual tasks and recommend evidence-based practices for the design of data graphics based on those findings [8], [9], [10].

We know less, however, about how humans translate that perception into real-world conclusions. In particular, we want to know how well the design of a data visualization impacts user understanding and interpretation of data graphics. Understanding the cognitive processes - beyond perception - that guide the user in making decisions or drawing conclusions about the data can better inform insights on design principles.

Digital Object Identifier 10.1109/XXX.0000.0000000

Models have been previously proposed for estimating chart comprehension [11], [12], with some specifically focused on bars and lines [13], [14]. Interpretation of line charts has previously been found to depend on complexity, with trend reversals adding to complexity more than the plotted number of data points for a very small data set [15]. However, many of the experiments underlying chart comprehension models were performed with small or biased sample sizes, with small groups of undergraduate students being a commonly studied population in this field (e.g., [13], [14], [15]).

Small sample sizes in many prior studies also mean we do not know as much about how understanding may differ across groups. While there is evidence to suggest that graph literacy - or the individual's abilty to accurately understand a graphic - varies by level of education and socieconomic status [16], [17], these studies focus primarily on univariate displays and simple graphics. Galesic and Garcia-Retamero developed and tested a scale for graph literacy among adult populations in the U.S. and Germany using probabilistic sampling [18]. The Galesic and Garcia-Retamero study focused on simple data visualization forms and comprehension tasks and posited that basic graph comprehension relies on some level of formal education or exposure to charts.

An improved understanding of viewer interpretation and how it differs across population groups will aid in the creation of data visualizations that communicate the intended information effectively. To this end, our study focuses on measuring to what extent U.S. adults understand the content in data graphics, whether the structure of the visualization impacts that understanding, and how this understanding differs across demographic groups. We fielded a survey-based user study with a nationally-representative, probability-based sample of U.S. adults. In this study, we vary the structure of charts presented and ask participants to answer questions about the content therein, focusing on their understanding of the data shown and the real-world conclusions drawn from the data.

The remainder of this paper is structured as follows: first, we describe the study design, including details on our sample of participants and the design of experimental images and questions. Secondly, we summarise the resulting participant sample and discuss participant behavior during the experiment. Subsequently, we discuss participant responses to our study questions, including their response patterns, accuracy of their responses, and how accuracy differs across demographic groups and other participant characteristics. Finally, we summarise our analyses and findings and discuss the implications of this work for the data visualization researcher and practitioner communities.

Study Design

We leveraged a survey format to conduct our user study, presenting survey participants with chart images and asking them to answer survey items about the content of the charts. The study has two major design components: the survey sampling and weighting strategy, and the design of the stimuli and corresponding survey questions.

Sample Design

We utilized NORC's AmeriSpeak Omnibus panel, a biweekly survey that samples from a standing panel of over 54,000 members aged 13 and over [19] and results in around 1,000 respondents in each round. The panel samples U.S. households using NORC's National Sample Frame, which provides coverage of over 97% of U.S. households. Due to the visual nature of our study, we excluded panelists responding through phone interviews, including only web-based panelists. Web-based respondents make up 93.3% of the total panel and 96.3% of the total panel weights.

Sampling was conducted across 48 sampling strata – split by age, education, gender, and race and Hispanic ethnicity. Each sampling stratum's size was determined by the corresponding population distribution, and takes expected differential survey completion rates into account in order to achieve a representative sample of the target population – here, U.S. adults. Resulting survey data were weighted to the U.S. Census Bureau's Current Population Survey (CPS) benchmarks and balanced by gender, age, education, race/ethnicity, and geographic region.

Experimental Design

Using data on the distribution of living arrangements among older age-sex groups in the United States, we created three different visual stimuli - a stacked bar chart, a diverging stacked bar chart, and a line chart, pictured in Figure 1. The data presented, size of the chart, color scheme, and legend were identical across all three chart types, with the only difference being the structural design of the chart. Keeping the survey items and context identical across rounds allows us to assume that differences in responses are due to the structure and type of the chart. The displayed data – survey results on living arrangements of older age-sex groups – was selected to be a minimally

controversial topic. At the same time, the topic was also intentionally chosen to be straightforward and relevant to participants' lives. The stimulus presented here is of medium complexity, as complexity is assessed by [13].

In each round of the Omnibus, participants were shown one of the stimuli and questioned about the chart and data presented. First, participants were asked to determine which of five statements in the survey were supported by the data in the chart. We presented three statements that were supported by the chart (i.e., were true) and two statements that were not supported by the chart (i.e., were false). The default selection for participants was no selection of the item, which corresponds to an indication of false for each statement. Participants had to select the true statements. We refer to these items as "interpretation statement" items.

Which of the following statements is supported by the chart?

1) More than half of all 70-79 year old men live in households with others (TRUE).
2) In each age-sex category, more adults live in households than in nursing homes or other group quarters (TRUE).
3) Compared to men, a higher percentage of women in each age group live in households with others (FALSE).
4) The percentage of adults living in nursing homes or other group quarters increases with age (TRUE).
5) The percentage of men who live alone in a household decreases with age (FALSE).

The statements presented to participants vary in difficulty, reflecting the complexity of the tasks required to answer them correctly. For example, the first item is relatively simple, requiring a participant to identify a single visual element (percentage of 70-79 year old men living in households with others) and determine whether it is larger than 50%. This task involves relatively few visual elements and participants may not have to complete any comparisons to other elements to determine it is a true statement. Other items are more difficult: in item 2, participants have to identify which visual elements correspond to living in households (blue bars or lines/points), which correspond to living in nursing homes or other group quarters (orange bars or lines/points), and then compare those visual elements across each age-sex group. Determining whether the second item is supported by the chart requires a much more involved set of steps and comparisons than item 1. In the presentation of results, we sort items by their difficulty based on the number of visual elements required to assess each provided statement.

Following the completion of interpretation statement items, participants were given four items that required them to identify and estimate specific data values within the chart by providing an integer value between 0 and 100.

What is the approximate size of the following? Just your best guess is fine.

1) Among 80-89 year old men, what is the percentage living with others in households?
2) Among 100+ year old women, what is the percentage living in a nursing home?
3) Among 70-79 year old men, what is the percentage living alone in a household?
4) Among 90-99 year old women, what is the percentage living alone in a household?

For each of these items, participants must identify the correct visual element within the chart (e.g., the element representing the percentage of 80-89 year old men living with others in households) and subsequently estimate the corresponding value using either the size (stacked bar and diverging stacked bar) or position (line chart) of that visual element. While estimating these values is primarily a perception task, it also requires identifying the correct chart element with provided context. This tests users' ability to understand legends and labels and apply that information to the real-world data in the chart. Each of these items are assumed to carry a similar level of difficulty as the required steps are identical with the exception of the final step, estimating the value itself. These items should also carry a similar difficulty to the first interpretation statement item. We will henceforth refer to this set of items as the "value estimation" items.

In designing the study items, we hypothesized that the interpretation statements carry a higher degree of difficulty than the value estimation questions. A basic understanding of the context within the chart is required to identify the corresponding elements in both sets of items; however, the interpretation statements require further inspection of the relationships between the represented values.

Participants were first shown all five interpretation statement items on a single screen with the accompanying chart. After completing those, participants were shown all four value estimation questions on a screen, again with the corresponding chart image. Participants could not go backwards to the interpretation statements to change their response. On each screen, participants could choose to zoom in to a larger version of the chart.

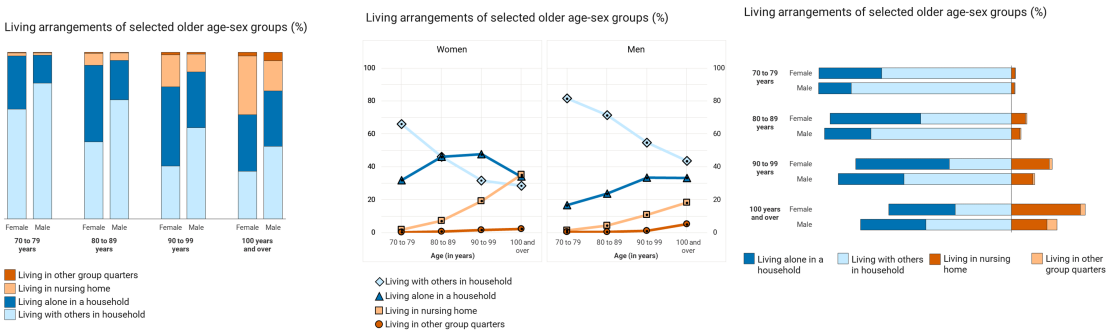

FIGURE 1. The visual stimuli shown to participants in the experiment. Participants saw either a stacked bar (left), line chart (center), or diverging stacked bar (right) representation of the same data in each round.

Participants

A total of 3,176 panelists participated across the three rounds. The resulting data from each of the three rounds were combined into a shared response file. Their corresponding sample weights were rescaled with respect to the total response population across all three rounds in order to preserve their respective weights following the combination method described by O'Muircheartaigh and Pedlow [20]. Indicators for round of origin were appended in the combined file. The number of responses and corresponding effective sample sizes and weights for each round of the study are shown in Table 1.

The distributions of self-reported demographic characteristics of participants across the rounds are shown in Table 2. The percentages are calculated marginally using survey weights. The distribution of demographic characteristics are substantively similar across the three rounds of this study. The "Asian-Pacific Islander, non-Hispanic", "2+, non-Hispanic", and "Other, non-Hispanic" race/ethnicity groups were grouped into a shared "All others, non-Hispanic" for our analyses due to small sample sizes in the distinct groups.

Participant Behavior

Metadata on participant behavior was captured, including device used to complete the survey and time spent on each set of questions, and the use of zooming. We determined that participants who completed the interpretation statement items in less than 25 seconds or the value estimation items in less than 20 seconds – an assumption of spending less than or equal to 5 seconds per item in the set – were speeding through the questions. We assume it would be very difficult to read an item, determine an answer, and provide that answer within five seconds or less [21]. It is possible that some of these identified speeders meaningfully completed several sub-items before dropping off due to fatigue or frustration. We did not remove these participants outright but rather appended an identifier for speeding behavior and incorporated it as a predictor in all modeling, so we could better understand how speeders' accuracy differs on our sets of items.

Notably, speeding behavior was more prevalent on the interpretation statement items (n = 686) as compared to the value estimation items (n = 141). This is interesting on multiple fronts: for one, since the interpretation statement items were presented first; if speeding behavior was due to survey fatigue alone, we would expect to see similar or higher rates of speeding in the latter set of questions. We observed the reverse, with participants more likely to speed through the first set of questions. We also assumed the interpretation statements should take more time per item, as they require more visual elements to address the validity of the statement; however, we see more respondents completing these items quickly. This could be attributed to frustration with the increased difficulty of the interpretation statement items or a lack of understanding of how to answer the questions. This is supported by the demographic patterns we observe among speeders; those with lower levels of educational attainment and lower income levels had higher rates of speeding behavior (see Figure 2).

One additional contributing factor to the differential speeding rates could be that the interpretation statements were presented in a 'select all that apply' format which requires clicking on the relevant corresponding items within the set, whereas the value estimation items require an input value separately for each item. Setting aside the content of the items, the minimal

TABLE 1. Rounds of study: Number of respondents (sample size), effective sample size, sum of weights and estimated lambda value factors used for combining surveys weights across the three rounds of the study.

Design	# Respondents	Effective Sample Size	Sum of Weights	Lambda values
Stacked Bar	1109	553.0	578.6	0.335
Diverging Stacked Bar	1012	551.6	526.7	0.334
Line	1055	545.0	542.4	0.330

TABLE 2. Self-reported participant demographic characteristics. Percent of weighted sample in each demographic group shown. Demographic representation by Age, Education, Gender, Income, and Race/Ethnicity is similar across the three distinct stimuli shown to participants. Values within a given chart type and demographic may not sum to 100 due to rounding.

Demographic	Group	Stacked Bar	Diverging Stacked Bar	Line Chart
Age	18-29	19.6	20.2	19.9
	30-44	25.6	26.8	25.3
	45-49	24.7	22.4	24.1
	60+	30.0	30.6	30.7
Education	Less than HS	9.7	9.4	9.0
	HS graduate or equivalent	27.6	28.9	28.8
	Some college/associates degree	27.1	26.4	26.5
	Bachelors degree	21.4	22.3	21.2
	Post graduate study/professional degree	14.1	13.1	14.5
Gender	Female	50.7	51.0	51.3
	Male	49.3	49.0	48.7
Income	Less than $30,000	20.8	20.5	21.9
	$30,000 to under $60,000	27.3	24.7	27.8
	$60,000 to under $100,000	25.1	23.3	21.8
	$100,000 or more	26.8	31.4	28.6
Race/Ethnicity	White, non-Hispanic	61.9	61.3	61.3
	Black, non-Hispanic	12.2	12.1	12.1
	Hispanic	17.0	17.5	17.5
	All others, non-Hispanic	9.0	9.1	9.1

time required to complete the value estimation items should be higher. Nevertheless, our assumptions on the time to read the statements, assess them, and reach a conclusion still hold; we simply acknowledge that it is physically easier to speed through the interpretation statements than the value estimation items if a participant is already going quickly.

Results

We present results on the response patterns and accuracy of participant responses for the interpretation statement and value estimation items. Subsequently, we discuss the relationship between responses across the two sets of items and the implications of that relationship.

Interpretation Statements

Each of the interpretation questions requires participants to complete a distinct task and utilize different sets of information in the chart. We first investigate the overall accuracy of our participants on each item by chart design, shown in Figure 3. We define correctness as a binary variable indicating whether the participant correctly selected the corresponding item (in the case of True statements) or did not select the corresponding item (in the case of False statements).

Items 5 and 3 have the highest levels of accuracy

FIGURE 2. Speeding behavior by participants. Participants were flagged as speeding on each set of items if they spent less than 5 seconds per item, on average, within a set (25 and 20 seconds for interpretation statement and value estimation items, respectively). Participants were more likely to speed through the interpretation statement items.

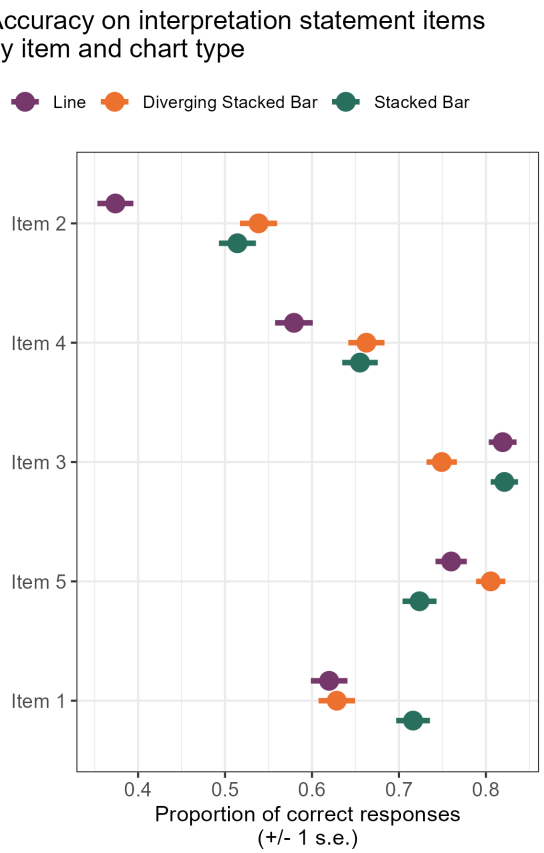

FIGURE 3. Percentage of respondents who selected the correct answer on each of the 5 interpretation statement items by chart structure. Items are ordered from most difficult (top, Item 2) to least difficult (bottom, Item 1) based on how many visual elements within the chart are required to assess the validity of the statement. Weighted survey mean and standard error bars are shown.

across all three chart structures. Notably, there are no clear patterns for any given chart design having the highest or lowest accuracy across all five items; the stacked bar excels for Item 1, put performs similarly to the diverging stacked bar in Items 4 and 2, and slightly lower in Item 5. The Line format appears to have lower performance for the most difficult items – Items 2 and 4, but does well in Items 3 and 5.

While we assume Item 1 is the least difficult item due to the fact that only one visual element is required to assess it, we also note that false statements (Items 3 and 5) are by nature easier than true statements when multiple visual elements are involved. True statements involving multiple visual elements must hold true across all those visual elements, while a false statement only has to be disproven by a single example or visual element to be determined to be false. We also note that True/False items should have a baseline 50% accuracy rate by nature if participants guess without considering the content.

Accuracy of responses We fit a logistic regression to model the proportion of correct responses out of 5 items, under chart type j for respondent k:

$$\text{logit } P(Y_{jk}) = \alpha_j + X_k\beta \quad (1)$$

where $j = \{1, 2, 3\}$ indicate chart type of the visual stimuli shown to participants, with 1 = Stacked Bar, 2 = Diverging Stacked Bar, 3 = Line, and assume participants $k = 1, \ldots, n$. Note that X_k represents the matrix of demographic characteristics of participant k, including 4-category Race/Ethnicity, 4-category age group, 4-category income group, 5-category education level, and 2-category gender.

The resulting model estimates are displayed in Table 3.

The stacked bar and diverging stacked bar charts resulted in the highest proportion of correct responses, on average, with the line chart corresponding to a minor, but significant, decrease in proportion of correct responses. Speeding behavior had a major impact on accuracy, unsurprisingly, with those who did not speed seeing significantly higher rates of accuracy than those who did speed. Those using a tablet, smartphone, or other non-desktop device to complete the web-based survey experienced a decrease in accuracy as well.

After accounting for differences across designs, speeding behavior, and device used, the proportion of correct responses increased significantly with higher levels of educational attainment. Gender and age were not significant predictors of accuracy after accounting for education, but older age groups had lower accuracy levels.

We also investigate item-specific logistic regression models to understand differences across different types of interpretation statement items. We utilize a logistic regression for correctness on item i under chart type j for respondent k and model:

$$\text{logit } P(Y_{ijk} = 1) = \mu_i + \alpha_j + X_k\beta \tag{2}$$

We fit a distinct model for each item i, simplifying the form to:

$$\text{logit } P(Y_{1jk} = 1) = \mu_1 + \alpha_j + X_k\beta \tag{3}$$

for Item 1, for example. The corresponding model coefficients for each of the 5 models are shown in Figure 4.

We observe that those who did not speed did significantly better on Items 1, 4, and 2; a similar pattern is observed for higher levels of educational attainment. Of note is that Items 1, 4, and 2 were the true items, requiring a respondent to explicitly select 'true' to get the item correct. This likely contributes to the significant increase in accuracy among non-speeders. The line chart design was associated with lower levels of accuracy on items 1, 4, and 2 while the diverging stacked bar was a mixed bag – resulting in increased accuracy on Item 5, but decreased accuracy for Items 1 and 3. Use of a tablet, smartphone, or other non-desktop device decreased accuracy on some items, but this was not universally true.

Value Estimation

We assume each value estimation question presents the same baseline difficulty to participants, as the task is structured similarly for all: participants must identify the relevant age group, sex group and living situation and subsequently estimate the value of the corresponding visual element shown in the chart.

Response patterns We first investigate the response patterns across each of the four value estimation items. The distribution of the raw difference from participants' answers to the true value is shown in Figure 5. While the median differences are similar across each chart design for most items, we note that the line chart median difference is closest to zero when summed across the four items. Additionally, on Item 2 the diverging stacked bar led to a positive median difference, while the stacked bar led to a slightly negative median difference. We observe the smallest IQRs under the line chart for Items 1, 2, and 3. Item 4 resulted in IQRs of similar magnitude across all three chart types. The diverging stacked bar design also led to a very large IQR for the raw difference from the true value on Item 1, which had a true value of 71.4.

Accuracy of responses To assess overall accuracy of participants' responses to the item set, we define correctness as a binary variable indicating whether the participants' response was within 5 points of the true value presented in the chart.

Similar to our model for accuracy on interpretation statement items, we fit a logistic regression to model the proportion of correct responses out of 4 items. We assume the same model format as Equations (2) and (3) and utilize the same set of participant behavior and demographic characteristics. The resulting model estimates are displayed in Table 4.

The line chart format resulted in the highest total rates of accuracy across the four value estimation items; while this is inconsistent with the interpretation statement results it is not particularly surprising since the line chart format is also the only format with a reference grid and percent labels along the relevant axis. Those who did not speed through the value estimation items had significantly higher levels of accuracy, while those who completed the survey on a tablet, smartphone, or other non-desktop device had significantly lower levels of accuracy.

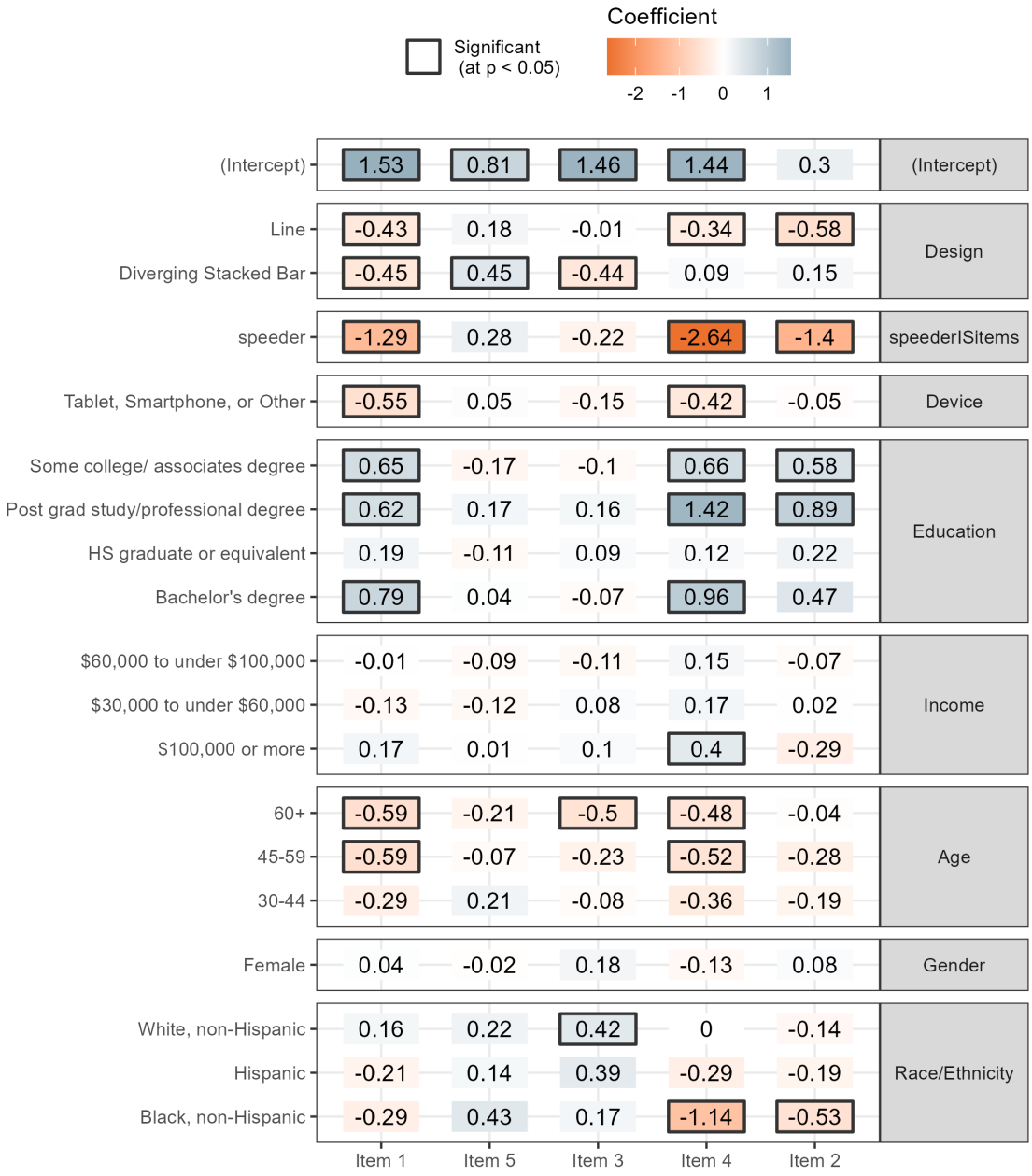

FIGURE 4. Model coefficient plot for logistic regression models predicting accuracy on each of the five interpretation statement items. The value shown in each square denotes the fitted model coefficient for that predictor. The backgound color denotes the size of the estimate, with darker blue values indicating a higher positive coefficient and darker orange indicating a larger negative coefficient. The grey border indicates significance of the model coefficient. Items are ordered by assumed level of difficulty, with Item 1 being the simplest and Item 2 being the most difficult.

TABLE 3. Model summary table for logistic model for proportion of correct responses to interpretation statement items. Log odds ratio (OR), 95% confidence interval (CI) for each estimate, and corresponding p-value are presented for each model predictor. The level of each variable represented in the model intercept is denoted by a blank row, where relevant.

Characteristic	log(OR)	95% CI	p-value
Design			
Stacked Bar	—	—	
Diverging Stacked Bar	-0.01	-0.04, 0.03	0.611
Line	-0.07	-0.10, -0.03	**<0.001**
speederISitems			
non-speeder	—	—	
speeder	-0.38	-0.43, -0.34	**<0.001**
Device			
Desktop	—	—	
Tablet, Smartphone, or Other	-0.06	-0.09, -0.03	**<0.001**
Education			
Less than HS	—	—	
HS graduate or equivalent	0.03	-0.05, 0.11	0.484
Some college/ associates degree	0.10	0.03, 0.17	**0.009**
Bachelor's degree	0.12	0.05, 0.20	**0.002**
Post grad study/professional degree	0.17	0.09, 0.25	**<0.001**
Income			
Less than $30,000	—	—	
$30,000 to under $60,000	0.00	-0.05, 0.05	0.952
$60,000 to under $100,000	-0.01	-0.06, 0.04	0.716
$100,000 or more	0.01	-0.04, 0.06	0.634
Age			
18-29	—	—	
30-44	-0.04	-0.08, 0.01	0.135
45-59	-0.09	-0.14, -0.04	**<0.001**
60+	-0.09	-0.14, -0.04	**<0.001**
Gender			
Male	—	—	
Male	—	—	
Female	0.01	-0.02, 0.04	0.507
Race/Ethnicity			
Other, non-Hispanic	—	—	
White, non-Hispanic	0.03	-0.02, 0.08	0.211
Black, non-Hispanic	-0.09	-0.16, -0.01	**0.021**
Hispanic	-0.01	-0.07, 0.04	0.635

Demographics play a major role in respondent accuracy; higher levels of educational attainment and income were associated with a higher number of correct responses, while increasing age was associated with lower numbers of correct responses. The Black non-Hispanic and Hispanic groups had significantly lower rates of accuracy than the Other non-Hispanic and White non-Hispanic groups. As in the interpretation statement model, gender was not a significant predictor of accuracy.

In contrast to the interpretation statement results, higher levels of income were associated with higher rates of accuracy on the value estimation items even after accounting for education.

Relationship between Item Sets

Finally, we investigate the relationship between participant accuracy on the two item sets: interpretation statements and value estimation. In Table 5 we display model coefficients for a logistic regression for the number of interpretation statements correctly answered. After accounting for speeding behavior, we find that each additional value estimation item correctly answered is associated with a significant increase in the proportion of interpretation statement items answered correctly. Greater ability to interpret the context within a chart, identify the correct element, and estimate its value, also indicates a stronger ability to accurately assess a set of real-world conclusion statements about the content shown in the chart.

On the contrary, in Table 6, we see that the same cannot be said for the ability to answer the interpre-

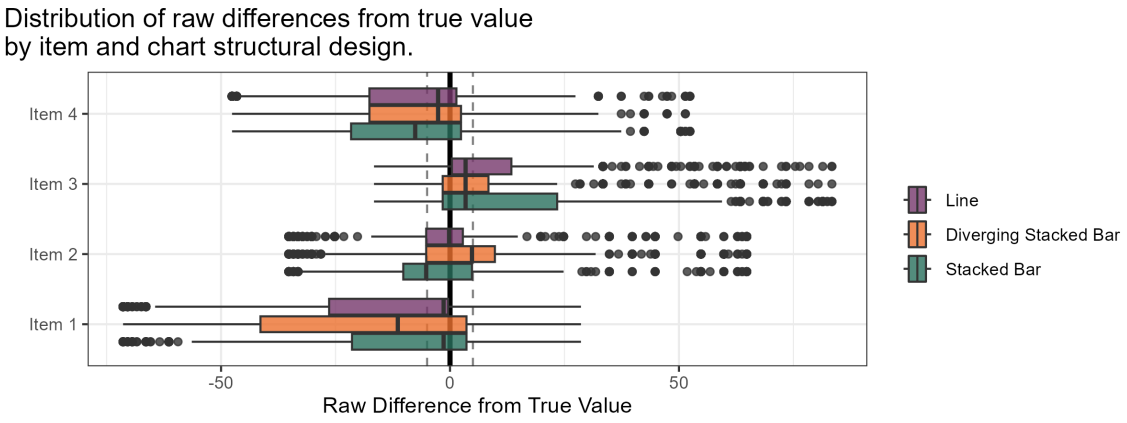

FIGURE 5. Distribution of participants' raw difference from the true value for each of the value estimation items shown in a boxplot. Vertical dashed grey lines are placed at -5 and 5. The line chart resulted in the minimal raw difference across the four items, although for each item one of the other designs performed at least as well as the line chart. Note that the range of raw differences shifts as it is the calculation of the value the respondent provided (a number 0 to 100) and the true value.

tation statement items. After accounting for speeding behavior, a higher number of interpretation statement items correct – in particular, 3 or more correct out of 5 – does not significantly increase ability to answer the value estimation questions.

These findings on the relationship between the two item sets are consistent with our design assumptions on the difficulty of the items; a greater ability to correctly complete the simpler value estimation items indicates a higher baseline understanding of how to read the chart and interpret its values, which in turn supports assessment of the validity of real-world conclusions about those values. Ability to assess real-world conclusions about the values does not, however, necessarily indicate a greater perceptual ability in the simpler tasks.

Supplementary Materials

Final survey data files and all code used in analyses and preparation of tables and figures in this paper are available online in a GitHub repository: https://github.com/kiegan/understanding-patterns-data-graphics.

Conclusions

Chart structure and context impacts how effectively viewers can interpret a data graphic. This includes their ability to identify and estimate values in context and assess the validity of real-world conclusions. The line chart format resulted in higher rates of accuracy for the value estimation items, but resulted in lower rates of accuracy overall for the interpretation statement items as compared to the stacked and diverging stacked bar formats. This suggests that while the line chart (with grid lines and corresponding axis labels) supports more accurate value recall after identifying the corresponding element within the chart, it may not universally be better for identifying all patterns or relationships mentioned in the interpretation statements. The bar chart formats may be more suitable to assessing the validity of more complex statements. There is an interplay between the goal of a chart (e.g., the conclusion a viewer might draw) and the selected format in determining the effectiveness of a given chart. Different formats can influence how the viewer interprets and extracts information about the same data.

Our results suggest a complex relationship between educational attainment and the interpretation of data graphics; U.S. adults with lower levels of educational attainment were more likely to speed through the more difficult interpretation statement questions, but rates of speeding were overall lower on the subsequent (simpler) value estimation questions, suggesting a level of frustration or discomfort with assessing the more complex interpretation statements. Further, after accounting for this speeding behavior, higher levels of educational attainment were associated with higher levels of accuracy on both interpretation statement and value estimation items.

Age also plays a role; an increase in age grouping was associated with lower levels of accuracy on both the interpretation statement and value estimation items.

TABLE 4. Model summary table for logistic model for proportion of correct responses to value estimation items, where a correct response is a value within 5 of the true value. Log odds ratio (OR), 95% confidence interval (CI) for each estimate, and corresponding p-value are presented for each model predictor. The level of each variable represented in the model intercept is denoted by a blank row, where relevant.

Characteristic	log(OR)	95% CI	p-value
Design			
Stacked Bar	—	—	
Diverging Stacked Bar	0.03	-0.05, 0.12	0.427
Line	0.28	0.20, 0.36	**<0.001**
speederVEitems			
non-speeder	—	—	
speeder	-1.4	-1.8, -0.91	**<0.001**
Device			
Desktop	—	—	
Tablet, Smartphone, or Other	-0.23	-0.30, -0.16	**<0.001**
Education			
Less than HS	—	—	
HS graduate or equivalent	-0.05	-0.23, 0.13	0.597
Some college/ associates degree	0.13	-0.03, 0.29	0.124
Bachelor's degree	0.22	0.06, 0.38	**0.008**
Post grad study/professional degree	0.27	0.10, 0.44	**0.001**
Income			
Less than $30,000	—	—	
$30,000 to under $60,000	0.16	0.03, 0.29	**0.014**
$60,000 to under $100,000	0.22	0.09, 0.35	**<0.001**
$100,000 or more	0.27	0.15, 0.39	**<0.001**
Age			
18-29	—	—	
30-44	-0.02	-0.13, 0.10	0.788
45-59	-0.13	-0.26, -0.01	**0.034**
60+	-0.24	-0.37, -0.12	**<0.001**
Gender			
Male	—	—	
Male	—	—	
Female	0.00	-0.07, 0.07	0.944
Race/Ethnicity			
Other, non-Hispanic	—	—	
White, non-Hispanic	0.06	-0.04, 0.17	0.223
Black, non-Hispanic	-0.26	-0.42, -0.09	**0.003**
Hispanic	-0.19	-0.33, -0.04	**0.012**

TABLE 5. Regression table for the number of interpretation statement items correctly answered by speeding behavior and number of value estimation items correctly answered. A higher number of value estimation items answered correctly is associated with an increased number of interpretation statement items correctly answered, after accounting for speeding behavior.

Characteristic	log(OR)	p-value
speederISitems		
non-speeder	—	
speeder	-0.32	**<0.001**
factor(numcorrectVE)		
0	—	
1	0.10	**<0.001**
2	0.20	**<0.001**
3	0.23	**<0.001**
4	0.27	**<0.001**

Our findings highlight the need for more comprehensive study on data visualization design and interpretation, and in particular, the barriers or biases to effective comprehension of visual displays of data that may exist across demographic groups within the U.S. Our study demonstrates basic differences across three structures – two bar formats and one line format – but focuses on a single data set, with a single data topic, and a single set of 9 distinct items (5 interpretation statement, 4 value estimation).

Observing the stark differences in interpretation ability and interaction with our study items that we do even within this limited scope invites further questions. For example, how might this differ with more drastic differences in chart design, data topics or data sets, or across other mapping structures (e.g., trellis plots,

TABLE 6. Regression table for the number of value estimation items correctly answered by speeding behavior and number of interpretation statements correctly answered. A higher number of interpretation statements answered correctly (i.e., 3 or more items) is not associated with an increased number of value estimation items correctly answered, after accounting for speeding behavior.

Characteristic	log(OR)	p-value
speederVEitems		
non-speeder	—	
speeder	-1.2	**<0.001**
factor(numcorrectIS)		
0	—	
1	-0.69	**<0.001**
2	-0.42	**0.030**
3	-0.27	0.131
4	0.18	0.322
5	0.27	0.125

scatter plots, pie charts)?

These are critical questions for the data visualization community to consider; visual communication of data to an audience relies on both the effective design of the visual and the ability of the audience member to understand and relate to the presented information. A well-designed chart whose values or relationships can be effectively *perceived* may still not be accessible to the audience if the format, design, or presentation is not well-understood by that audience or there is a high cognitive burden to interpreting the information. Difficulty in interpretation or a high cognitive burden may also lead some audiences to simply not engage meaningfully with a chart, as we observe with the high number of participants speeding through the interpretation statement items. Evidence from our study suggests that data visualization designers should consider several factors: (1) their target audience's existing comfort level with data graphics, (2) the intended message they would like users to conclude (easily), and (3) the structural design that supports their target audience drawing that specific conclusion. While further work is needed in the field to better understand the landscape surrounding factors (1) and (3), considering these factors thoughtfully when designing a data graphic will create a more accessible graphic that more effectively reaches the target audience.

Acknowledgment

We thank Edward Mulrow for his insight on experimental design and survey sample combination. We also thank NORC's AmeriSpeak Omnibus team for their support in fielding these rounds of data collection and providing the final weighted data files to our team.

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*, vol. 79, no. 387, pp. 531–554, Sep. 1984, doi: 10.1080/01621459.1984.10478080.

[2] J. Heer and M. Bostock, "Crowdsourcing graphical perception: Using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Atlanta Georgia USA: ACM; ACM, Apr. 2010, pp. 203–212. doi: 10.1145/1753326.1753357.

[3] B. M. Hughes, "Just noticeable differences in 2d and 3d bar charts: A psychophysical analysis of chart readability," *Perceptual and Motor Skills*, vol. 92, no. 2, 2, pp. 495–503, 2001, doi: 10.2466/pms.2001.92.2.495.

[4] M. Lu *et al.*, "Modeling Just Noticeable Differences in Charts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, 1, pp. 718–726, Jan. 2022, doi: 10.1109/TVCG.2021.3114874.

[5] K. Rice, H. Hofmann, N. du Toit, and E. Mulrow, "Testing Perceptual Accuracy in a U.S. General Population Survey Using Stacked Bar Charts," *Journal of Data Science*, vol. 22, no. 2, pp. 280–297, Mar. 2024, doi: 10.6339/24-JDS1121.

[6] M. Correll, E. Bertini, and S. Franconeri, "Truncating the Y-Axis: Threat or Menace?" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr. 2020, pp. 1–12. doi: 10.1145/3313831.3376222.

[7] J. M. Hofman, D. G. Goldstein, and J. Hullman, "How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, in CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12. doi: 10.1145/3313831.3376454.

[8] C. M. Carswell, "Choosing specifiers: An evaluation of the basic tasks model of graphical perception," *Human factors*, vol. 34, no. 5, 5, pp. 535–554, 1992, doi: 10.1177/001872089203400503.

[9] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The Science of Visual Data Communication: What Works," *Psychol Sci Public Interest*, vol. 22, no. 3, pp. 110–161, Dec. 2021, doi: 10.1177/15291006211051956.

[10] Z. Zeng and L. Battle, "A Review and Collation of Graphical Perception Knowledge for Visualization Recommendation," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–16. doi: 10.1145/3544548.3581349.

[11] S. Pinker, "A theory of graph comprehension," in *Artificial intelligence and the future of testing*, 1st ed., Psychology Press, 1990, pp. 73–126.

[12] T. Munzner, "A Nested Model for Visualization Design and Validation," *IEEE Trans. Visual. Comput. Graphics*, vol. 15, no. 6, pp. 921–928, Nov. 2009, doi: 10.1109/TVCG.2009.111.

[13] P. Shah and E. G. Freedman, "Bar and line graph comprehension: An interaction of top-down and bottom-up processes," *Topics in Cognitive Science*, vol. 3, no. 3, 3, pp. 560–578, 2011, doi: 10.1111/j.1756-8765.2009.01066.x.

[14] J. Zacks and B. Tversky, "Bars and lines: A study of graphic communication," *Mem Cogn*, vol. 27, no. 6, pp. 1073–1079, Nov. 1999, doi: 10.3758/BF03201236.

[15] C. M. Carswell, C. Emery, and A. M. Lonon, "Stimulus complexity and information integration in the spontaneous interpretations of line graphs," *Applied Cognitive Psychology*, vol. 7, no. 4, pp. 341–357, 1993, doi: 10.1002/acp.2350070407.

[16] M.-A. Durand, R. W. Yen, J. O'Malley, G. Elwyn, and J. Mancini, "Graph literacy matters: Examining the association between graph literacy, health literacy, and numeracy in a Medicaid eligible population," *PLoS ONE*, vol. 15, no. 11, p. e0241844, Nov. 2020, doi: 10.1371/journal.pone.0241844.

[17] R. Garcia-Retamero, D. Petrova, A. Feltz, and E. T. Cokely, "Measuring Graph Literacy," in *Oxford Research Encyclopedia of Communication*, Oxford University Press, 2017. doi: 10.1093/acrefore/9780190228613.013.302.

[18] M. Galesic and R. Garcia-Retamero, "Graph literacy: A cross-cultural comparison," *Med Decis Making*, vol. 31, no. 3, pp. 444–457, 2011, doi: 10.1177/0272989X10373805.

[19] J. M. Dennis, "Technical overview of the AmeriSpeak panel NORC's probability-based household panel," *NORC at the University of Chicago*, 2019, updated 2022.

[20] C. O'Muircheartaigh and S. Pedlow, "Combining Samples Vs. Cumulating Cases: A Comparison of Two Weighting Strategies in NLSY97," in *ASA Proceedings of the Joint Statistical Meetings*, 2002, pp. 2557–2562. Available: http://www.asasrms.org/Proceedings/y2002/Files/JSM2002-001082.pdf

[21] M. Brysbaert, "How many words do we read per minute? A review and meta-analysis of reading rate," *Journal of Memory and Language*, vol. 109, p. 104047, Dec. 2019, doi: 10.1016/j.jml.2019.104047.

Kiegan Rice is a Senior Statistician at NORC at the University of Chicago. Her research interests include data visualization design, interactive data visualization development, and computational reproducibility. Dr. Rice received a Ph.D. in Statistics from Iowa State University. She is a member of the American Statistical Association and the Data Visualization Society. Contact her at rice-kiegan@norc.org.

Sydney Bell is a Data Analyst at NORC at the University of Chicago. Her research interests include interactive data visualization development and statistical communication. Bell received a bachelor's degree in Mathematics from Carleton College. Contact her at bell-sydney@norc.org.

Taylor Wing is a Statistician at NORC at the University of Chicago. Her research interest includes interactive data visualization development, visualization design, and survey statistics. She received a master's degree in Statistics with a concentration in data analytic methods from the University of Virginia and is a member of the American Statistical Association. Contact her at wing-taylor@norc.org.

Heike Hofmann is a Professor in the Department of Statistics at the University of Nebraska-Lincoln. Her research interest include the development of methodology for observational data and data exploration with a focus on statistical graphics. She is a fellow of the American Statistical Association and a member of the International Statistical Institute. Contact her at hhofmann4@unl.edu.

Nola du Toit is a Senior Research Methodologist and Data Visualization Specialist at NORC. Her research interests include data visualization, accessibility, and equity in scientific communication. She received her PhD in Sociology from Bowling Green State University. Contact her at dutoit-nola@norc.org.