Supplement: Analysis and code for 'Testing Perceptual Accuracy in Charts using Surveys'

July 12, 2023

Abstract

1 Survey rounds

The data for this paper were collected in several rounds as part of the NORC Omnibus.

Table 1: Survey rounds: dates, number of participants (nominal sample size), effective sample size, and sum of weights.

Name	Date	# Participants	effective sample size	Sum of weights $\sum_i w_i$
Round 1	April 2022	933	521.1	934.9
Round 2	May 2022	953	485.7	953.4
Round 6	Sep 2022	450 [Split]	254.9	462.9
Round 7	Oct 2022	984	524.5	984

We are using a strategy of *combining* (rather than cumulating) surveys S_1 and S_2, as described in **?**, by multiplying weights in S_1 and S_2 by λ and $1-\lambda$, respectively.

$$\lambda = \frac{n_1/d_1}{n_1/d_1 + n_2/d_2},$$

where n_1 and n_2 are the nominal sample sizes and d_1 and d_2 are the design effects for the estimators. Instead of using design effects itself, d_1 and d_2 are estimated as

$$d_1 = 1 + CV(w_i \in S_1)^2 \quad \text{and} \quad d_2 = 1 + CV(w_i \in S_2)^2$$

CV is the coefficient of variation of the weights within each sample, and is estimated as **?**

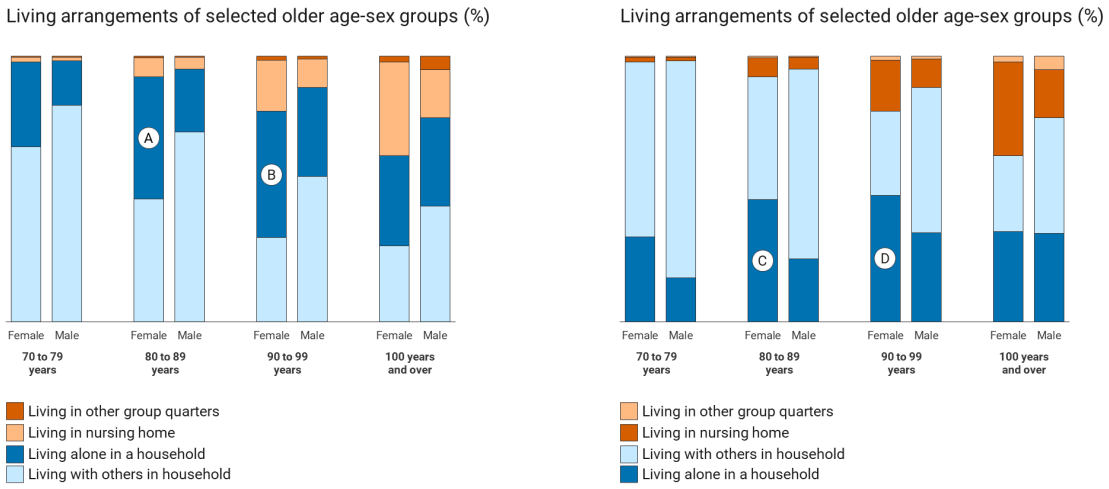
$$CV(w \in S) = \frac{\widehat{Var(w)}}{\bar{w}^2}.$$

? estimate λ separately for any combination of race/ethnicity by sex. We will use that strategy whenever we include demographic variables in the analysis, otherwise we will use a single adjustment for the weights.

All calculations are done in R (?) using the `survey` package (?) version 4.0 (?) based on ?.

2 Model 1: Comparing Aligned and Unaligned Tiles in Vertical Stacked Barcharts

The data used for this is a combination of rounds 1 and 2, with $\lambda = 0.518$ for an effective sample size of 1007.1. Figure ?? shows the two stacked barcharts shown to all panelists.



(a) **Unaligned tiles:** Is A bigger than B?

(b) **Aligned tiles:** Is C bigger than D?

Figure 1: The two stacked barcharts every participant got to see. In each barchart, the two marked tiles are to be compared for their size. in both instances, the tile on the right is (very slightly) larger.

2.1 Binary response

Defining an accurate response as "B is bigger" in the chart of unaligned tiles, and "D is bigger" in the aligned case, while calling the other two responses as incorrect, leads to the data shown in Figure **??**(a): we see that more than twice the number of responses is accurate, when the tiles are aligned along the same axis. Because each particpant was shown both versions of the chart, we can use a paired t-test to compare mean accuracy between the two charts. The difference in mean accuracy is 0.31. This difference is highly significant (t statistic: 15.94, df: 1656, p-value: $<$ 2.2e-16).

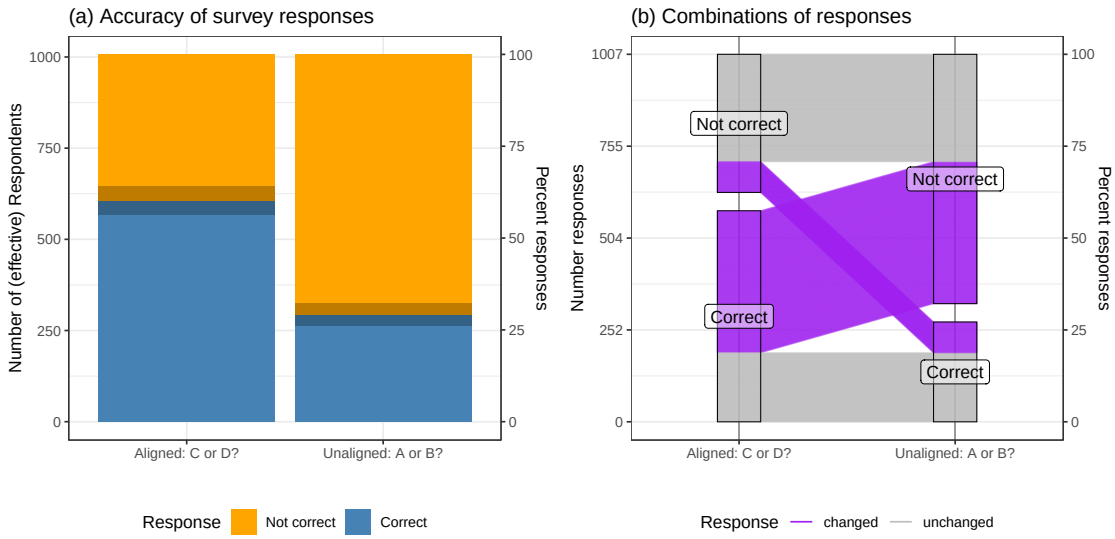


Figure 2: On the left (a), a stacked barchart shows the number of respondents with correct (blue) and incorrect (orange) responses to the two comparison questions. When tiles are aligned along the same axis, more than twice the number of responses is accurate. The shaded area along the top of the blue tiles corresponds to 95% confidence intervals around (marginal) correct responses. On the right (b), a parallel coordinate plot shows all combinations of responses. There's a huge asymmetry in the number of responses, where participants answered only one of the questions correctly. A lot more responses are correct when comparing aligned tiles than unaligned tiles.

This test is equivalent to a logistic regression with response Y_{ij}, the response of participant i to question Q_j $j = 1/2$ is correct (=1) or incorrect (=0):

$$\text{logit } P\left(Y = 1 \mid Q_j\right) = \alpha + \beta_j,$$

We assume that $Y \mid Q_j$ has a Bernoulli distribution with success probability p_j, $j = 1, 2$. For the purpose of estimability, we will assume that $\beta_1 = 0$, i.e. $\text{logit}(p_1) = \alpha$ and $\beta_2 = \text{logit}(p_2) - \text{logit}(p_1)$. The odds of an accurate answer with aligned bars is therefore exp $(1.3) = 3.7$ times higher than for the unaligned bars of Question 1.

Estimates for Model 1

accuracy of responses for unaligned (Question 1) and aligned bars (Question 2)

term	estimate	SE	t-statistic	p-value
$$\widehat{\alpha}$$	−0.89	0.07	−13.0	< 0.0001
$$\widehat{\beta}_2$$	1.30	0.09	14.9	< 0.0001

2.2 Ordinal response

When moving beyond the binary accuracy measure of the response and using all three levels of the response as dependent variable, we see that by far the largest change in responses comes from a change from 'they are the same' to the (perceptually) correct response of 'D is bigger' when the tiles are aligned along the same axis for a comparison (see Figure **??**). Even though the number of wrong responses (orange) is only slightly different (0.045) between aligned and unaligned comparisons, it is still highly significant (t statistic: 3.9, df: 1656, p-value: < 2.2e-16).

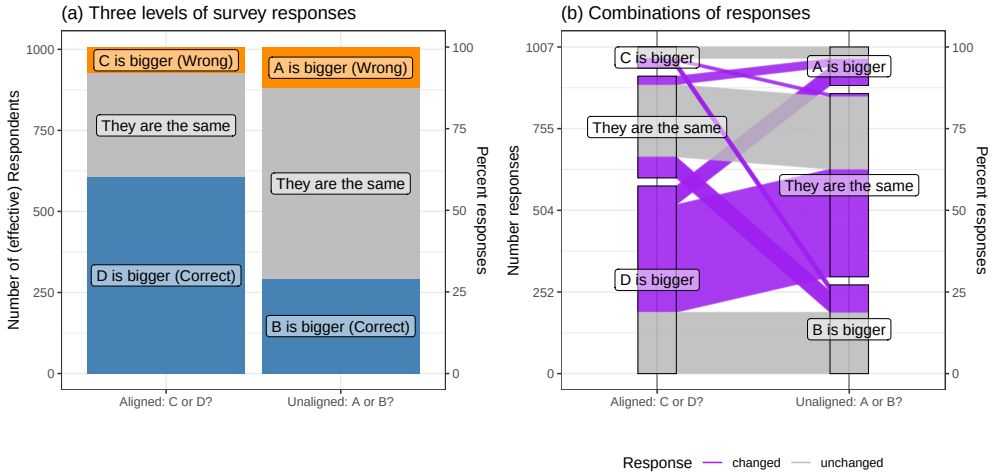
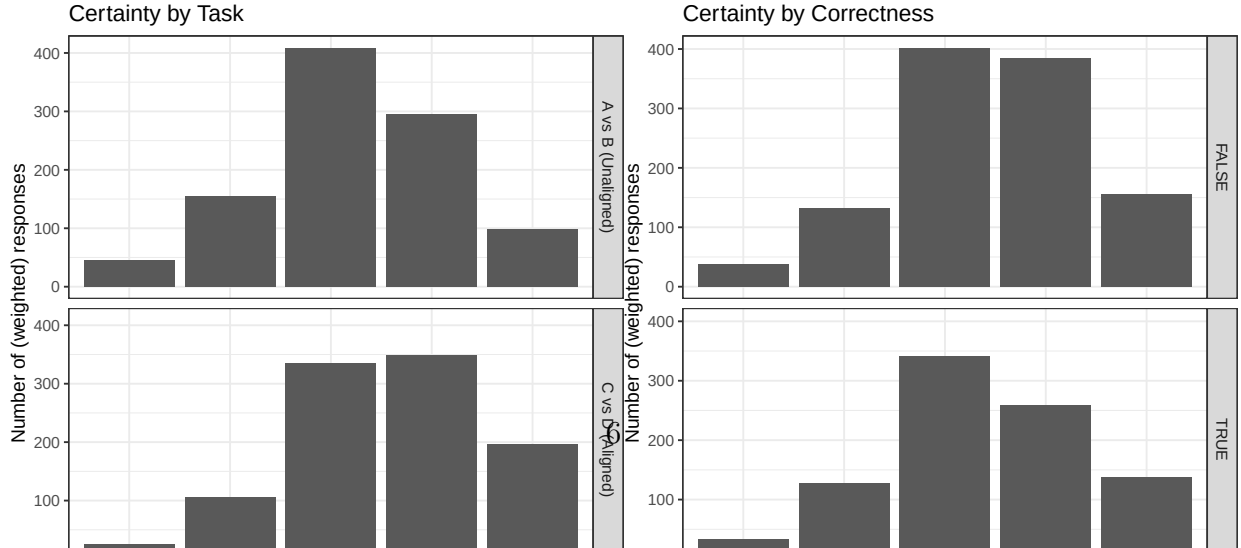


Figure 3: On the left (a), a stacked barchart shows the number of respondents for each level of response: by aligning tiles, the number of responses of 'they are the same' is cut in half. On the right (b), a parallel coordinate plot shows all combinations of responses. The largest nuber of changes are from 'D is bigger' to 'they are the same'.

Table 2: Table 2: A change from aligned to unaligned tiles is detrimental to the level of correct (or correct and same) responses.

level	log odds	SE	t-statistic	p-value	percent change
correct | same or wrong	-1.3	0.09	-14.9	0e+00	-31.0
correct or same | wrong	-0.5	0.13	-3.8	2e-04	-4.5

2.3 How certain are participants?



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 and correctness on certainty by using a cell-means model of the form:

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

where $k = 1, ..., N$, $\mu_{ij(k)}$ is average certainty (measured on a scale from 1 to 5) of the four combinations of task and correctness, where $i = 1, 2$ encodes unaligned/aligned, and $j = 1, 2$ encodes wrong, correct, respectively. We also assume that errors are normally distributed, i.e. $\epsilon_k \overset{i.i.d}{\sim} N(0, \sigma^2)$ for all $k = 1, ..., N$. What we find, is that we see the lowest score of certainty for correct responses to the unaligned comparison task. For aligned comparisons, the correctness of the response does not matter for the associated certainty.

Estimates for Model 2.3

certainty of responses by task and correctness of response

term	estimate	SE	t-statistic	p-value
$$\widehat{\mu}_{11}$$	3.34	0.04	84.5	< 0.0001
$$\widehat{\mu}_{21}$$	3.61	0.05	73.0	< 0.0001
$$\widehat{\mu}_{12}$$	3.01	0.06	52.4	< 0.0001
$$\widehat{\mu}_{22}$$	3.56	0.04	85.4	< 0.0001

```
Warning: Combining variables of class <logical> and <character> was deprecated in
ggplot2 3.4.0.
i Please ensure your variables are compatible before plotting (location:
  `combine_vars()`)
```

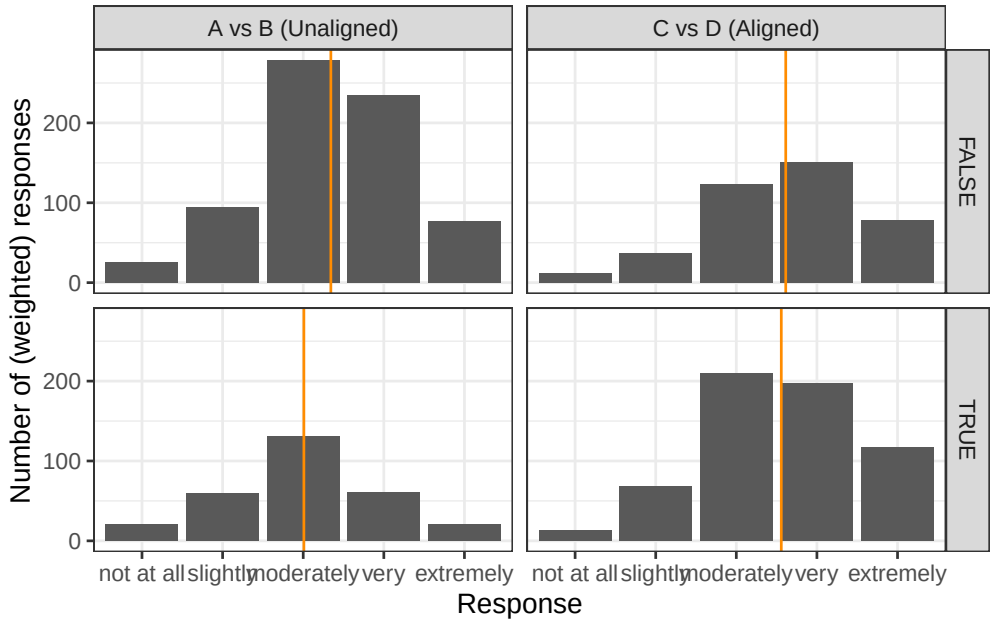


Figure 5: Certainty by task and correctness. Correct answers in the unaligned task have the least certainty associated with them. An incoorect response on the unaligned task shows a significant boost in certainty.

3 Model 2: Aligned and unaligned comparisons between facetted barcharts and facetted piecharts

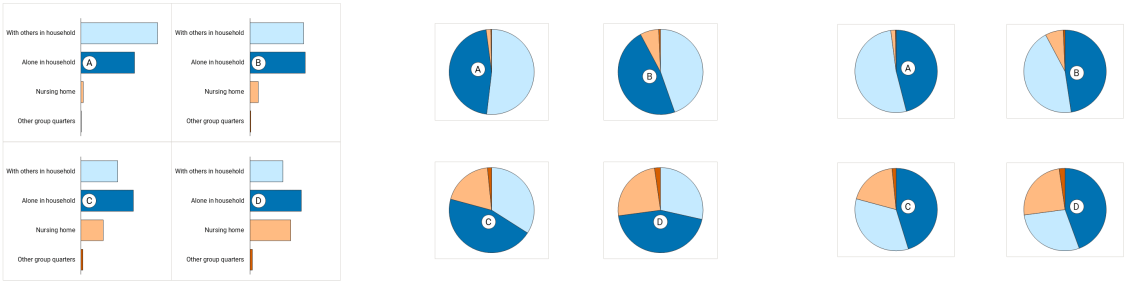
The data this section is based on comes from (a part) of Round 6 and all responses from Round 7.

A total of three different charts were shown to (a part) of the panelists from Round 6 and all panelists of Round 7. Each of the charts has four labelled pieces A, B, C, and D.

The questions asked for each of these charts were of the form: "which of the marked pieces is bigger? Just your best guess is fine".

A. A or B?

RESPONSE OPTIONS:



(a) **Round 6**: Facetted bar charts

(b) **Round 7**: Facetted piecharts with unaligned wedges

(c) **Round 7**: Facetted piecharts with aligned wedges

Figure 6: Stimuli shown to panelists. For each of the charts, all marked pieces were evaluated pairwise for their size difference.

1. A is bigger
2. B is bigger
3. They are the same

Each participant was asked a total of six comparisons of this type: A vs B, A vs C, A vs D, B vs C, B vs C, and C vs D. The number of valid responses and the resulting effective sample sizes are shown in Table **??**

```
Warning: HTML tags found, and they will be removed.
* Set `options(gt.html_tag_check = FALSE)` to disable this check.
HTML tags found, and they will be removed.
* Set `options(gt.html_tag_check = FALSE)` to disable this check.
```

Table 3: Table of the number of effective responses by task

Number of (effective) responses by task

task	number of responses:nominal	number of responses:effective	number of participants
Round 6			
Bar	2,618	1,480.3	447
Round 7			
Pie-Aligned	5,749	3,058.5	972
Pie-Unaligned	5,759	3,058.2	978

The four labelled pieces A, B, C, and D correspond to values 46, 47.6, 45.2, and 44.4, respectively, i.e underlying each evaluation is the order D < C < A < B. Nominally, the differences between the size of the marked elements can be expressed as multiples in $\delta = 0.8$, with $B = A + 2\delta = C + 3\delta = D + 4\delta$. A difference of $\delta = 0.8$ corresponds to about a 3 degree difference for the angle of a wedge. We would expect that an increase in the difference in the values increases the accuracy in the responses. This is mitigated by the distance between wedges: more distance (should) lead to a decrease in accuracy. Similarly, if pieces have an axis or a line of reference in common, we would expect an increase in accuracy.

3.1 Binary and Ordinal Correctness as Response

Figure **??** gives an overview of all responses to size assessments of the marked tiles. The panels are ordered according to the difference in signal and hypothesized difficulty from easiest (left) to hardest (right) comparison. We see, that in general the percentage of correct responses decreases from left to right. Design does not seem to matter for easy

comparisons (BD, BC) or the hard comparison (CD). The three comparisons in the middle show interesting patterns, highlighted in the reordered Figure **??**. Colored letters below and above the bars encode significances between pairs of bars of the same color: two proportions are significantly different at a 5% significance level, if they do not have a letter in common (**?**). Significances are adjusted for multiple comparisons as implemented in the `multcomp` package (**?**).

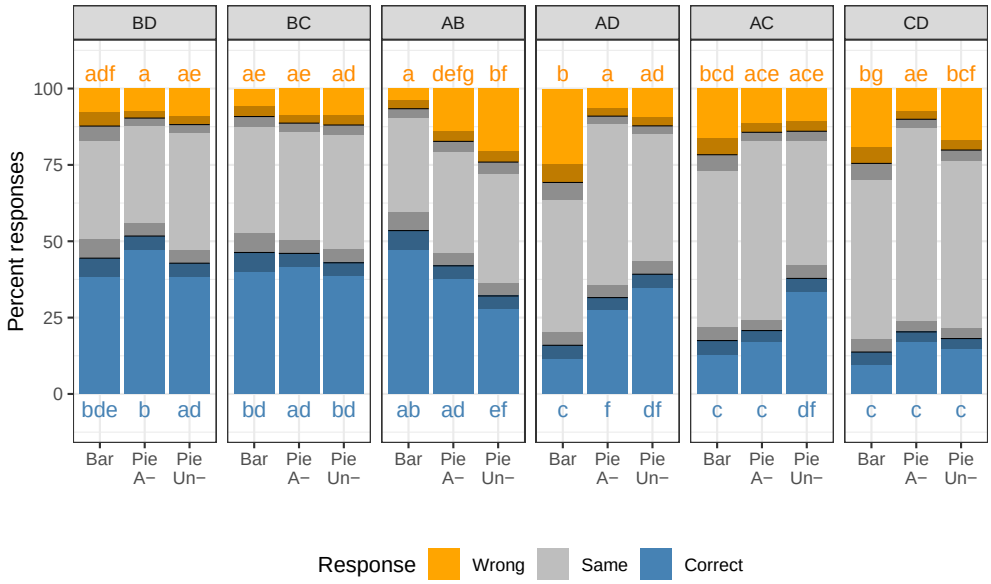


Figure 7: Responses for comparisons AB through CD on the three different facetted designs. Responses to the same comparison are shown side-by-side for the three designs. Comparisons are ordered from least difficult (BD) to most difficult (CD). The increasing difficulty of comparisons is reflected in the overall decreasing percentage of correct responses (height of blue tiles). The overlaid rectangles represent 95% confidence intervals. The letters in blue and orange encode significances between pairwise proportions: two bars have a significantly different proportion (at a 5% significance level) if they do not share a letter.

For the facetteed barcharts, there is an abrupt drop in accuracy of responses between 'easy' and 'hard' comparisons. (HH: All of these comparisons could be affected by the context of

their light blue neighbors. 'B' is bigger than its light blue neighbor, and also the overall biggest. 'A' is only a bit smaller than 'B', but quite a bit smaller than the light blue bar next to it. This seems to make comparisons with A harder.) For pie charts the accuracy in responses is reacting a lot more gentle to the increasing difficulty. There is no apparent effect in pie charts due to alignment. However, all comparisons are very close to 50% - the intrinsic reference lines (**?**) in pie charts (at 180 degrees, and to a lesser degree at 90 degree angles) might be helping with this particular set of comparisons.

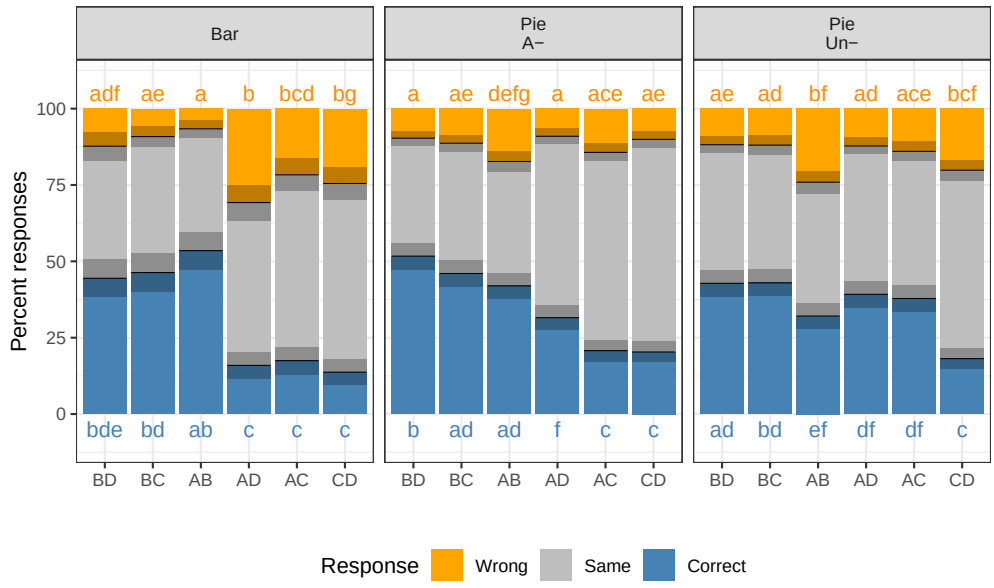


Figure 8: Responses for comparisons AB through CD on the three different facetted designs, reordered. Now, responses to the same design are shown side-by-side across comparisons. Comparisons are ordered from least difficult (BD) to most difficult (CD). The three different designs result in different response patterns for increasing comparison difficulty.

Besides intrinsic reference lines, the overall size of the stimulus might be another factor contributing to the difference in accuracy between the tiles and the wedges. In the shown example, the marked tiles have an area between $133 \times 55 = 7315$ pixels (for D) to $144 \times 55 = 7920$ pixels (for B), while the pie wedges are based on circles with a diameter of 105 pixels,

and therefore cover visually more than twice the area of the tiles in the barchart. The area of D in form of a pie-wedge is $\frac{44.4}{100} \times 105^2\pi \approx 15{,}378$ pixels and B has an area of $\frac{47.6}{100} \times 105^2\pi \approx 16{,}487$ pixels.

There are different possible strategies for evaluating the size of a wedge in a pie chart. Panelists can use the **angle** of the wedge (or a suitable derivative of it, such as the complementary angle or the remainder to 180 degrees), its **area** (as part of the circle or some derivative) or the **arc length** of the wedge. Studies by **?** and **?** show that the area of a piece in a pie chart is the primary evaluation method.

3.2 How much time did participants spend on each task?

Note: we do not have certainty for round 7. We also do not have timing by individual comparisons.

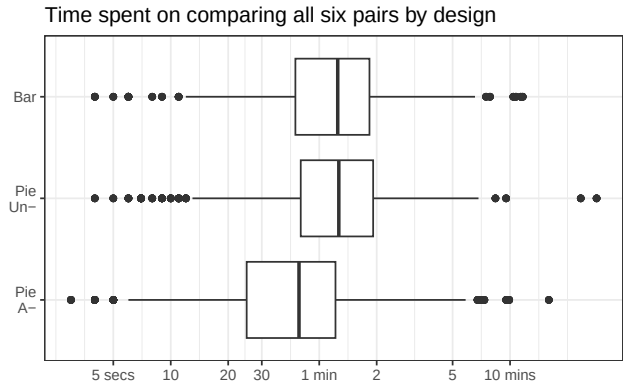


Figure 9: Comparisons for aligned pie slices were on average the fastest. Comparisons in the facetted barcharts took the longest. The difference in average times are significant for aligned pies versus either of the other designs, but not different between barcharts and unaligned pie charts.

```
Call:
```

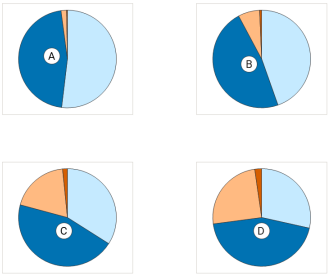
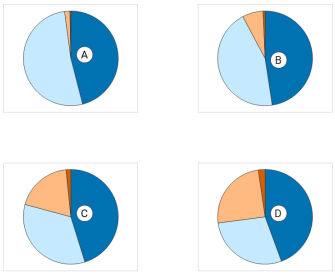
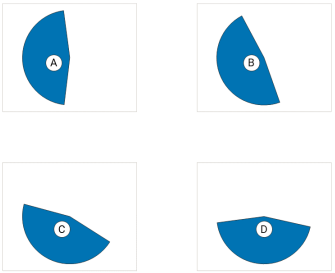
```
svyglm(formula = time_spent ~ task, design = survey_67)


Survey design:

svydesign(~CaseId, data = response_67, weights = ~weight)


Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)    98.841      5.865  16.851  < 2e-16 ***

taskPie\nA-   -40.637      6.267  -6.484 1.25e-10 ***

taskPie\nUn-   -6.351      6.712  -0.946    0.344

---

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


(Dispersion parameter for gaussian family taken to be 6933.628)


Number of Fisher Scoring iterations: 2
```

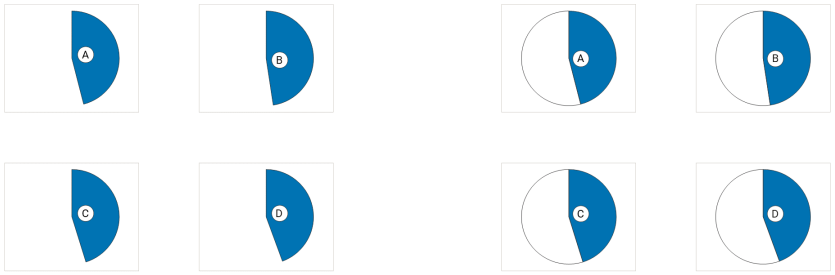
4 Model X: Comparing Aligned and Unaligned Elements Facets

The survey setup and all stimuli are shown in Table **??**. A total of 984 panelists (Effective sample size: 524.5) were shown four sets of facetted charts each. For questions 3 and 4, the panel was split into two (roughly) halves. One set of panelists was shown the floating wedges, while the other half was shown the framed wedges on the right.

For each chart, panelists were asked to compare all marked tiles against each other.

Table 4: Setup of questions and stimuli for Round 7.

Everybody:	**Question 1:** Facetted Pie, unaligned	**Question 2:** Facetted Pie, aligned
		
Split Sample	Effective sample size: 257.3	Effective sample size: 267.7
Question 3: unaligned	Floating pie wedges	Framed pie wedges
		



Question 4:

aligned

For a comparison of wedges marked X and Y, there were three possible options for the response: (1) X is bigger, (2) Y is bigger or (3) They are the same.

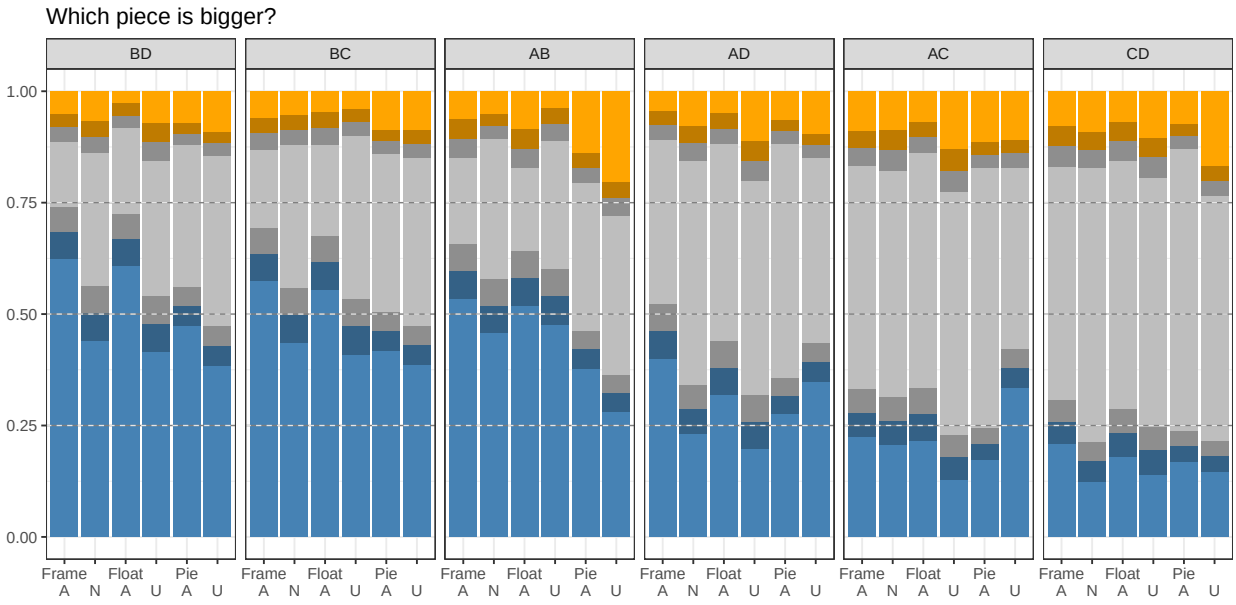


Figure 10: Panelist responses by task and comparison. On average, the accuracy for aligned pieces is higher than for unaligned pieces. Similarly, the accuracy for framed pieces is higher than for floating pieces, which in turn has a higher accuracy than the full pie chart.

Figure **??** shows a summary of panelists' accuracy of their responses. The panels are sorted according to the level of difficulty of a comparison as hypothesized above from lowest