E1 Results

Alex Kale

2021-03-02

Results

In this document, we present the results of our first experiement. This is intended as a supplement to the full paper. Here, we gather information from our exploratory analysis and model expansion processes, with an emphasis on what we've learned and an eye toward what should be presented in the paper.

Analysis overview

Our primary research questions focus on how visualization conditions impact the correspondence between user responses and our normative benchmark *causal support*. Causal support says how much a chart user should believe in alternative causal explanations given a data set. In our first experiment, we ask users to differentiate between two alternative causal models, one with a treatment effect (explanation A) and one without a treatment effect (explanation B).

We estimate the correspondence between user responses and our normative benchmark using a linear in log odds (LLO) model, where ideal performance is a one-to-one relationship between a user's responses and normative causal support. We chartacterize performance primarily in terms of LLO slopes with respect to causal support, which is a measure of sensitivity to the signal in each data set that should support causal inferences. The LLO also has an intercept term, which measures the average response when there is no signal to support either causal explanation. Intercepts represent an overall bias in responses to the extent that they deviate from 50%.

We look at how LLO slopes and intercepts very as a function of visualization condition. In interactive visualization conditions, we separate trials depending on whether or not users interacted with the visualization, which reduces statistical power in these conditions but enables us to be more accurate in estimating the effects of interactive visualizations on causal inferences. At the end of the document we follow up and do a descriptive analysis of how chart users interacted with these visualizations, specifically, which views of the data they chose to create.

Prepare data

Load data.

```
df <- read_csv("e1-anonymous.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   workerId = col_character(),
##   condition = col_character(),
##   gender = col_character(),
##   age = col_character(),
##   education = col_character(),
##   chart_use = col_character(),
##   problems = col_character(),
##   interactions = col_character(),
##   trial = col_character(),
##   userInput = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(df)
```

```
## # A tibble: 6 x 34
##   workerId batch bonus condition duration n_data_conds n_trials gender age  
##   <chr>    <dbl> <dbl> <chr>        <dbl>        <dbl>    <dbl> <chr>  <chr>
## 1 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## 2 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## 3 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## 4 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## 5 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## 6 b888e39c    29  0.25 text          567.           16       18 woman  55-64
## # … with 25 more variables: education <chr>, chart_use <chr>, problems <chr>,
## #   abs_err <dbl>, causal_support <dbl>, count_GT <dbl>, count_GnT <dbl>,
## #   count_nGT <dbl>, count_nGnT <dbl>, ground_truth <dbl>, interactions <chr>,
## #   n <dbl>, p_treat <dbl>, payoff <dbl>, q_idx <dbl>, response_A <dbl>,
## #   response_B <dbl>, total_GT <dbl>, total_GnT <dbl>, total_nGT <dbl>,
## #   total_nGnT <dbl>, trial <chr>, trial_dur <dbl>, trial_idx <dbl>,
## #   userInput <chr>
```

Calculate a log response ratio `lrr` to model as a function of `causal_support`. Also, convert predictor variables to factors for modeling if need be.

```
model_df <- df %>%
  # drop practice trial
  filter(trial != "practice") %>%
  mutate(
    # response units
    response_A =  if_else(
      response_A > 99.5, 99.5,
      if_else(
        response_A < 0.5, 0.5,
        as.numeric(response_A))),
    response_B =  if_else(
      response_B > 99.5, 99.5,
      if_else(
        response_B < 0.5, 0.5,
        as.numeric(response_B))),
    lrr = log(response_A / 100) - log(response_B / 100),
    # predictors as factors
    worker = as.factor(workerId),
    vis = as.factor(condition),
    n = as.factor(n),
    # derived predictors
    delta_p = (count_nGnT + count_GnT)/(total_nGnT + total_GnT) - (count_nGT + count_G
T)/(total_nGT + total_GT),
    interactions_processed = if_else(interactions == "placeholder", list(NA), str_split
(interactions, "_")),
    trial = as.numeric(trial),
    trial_n = (trial - mean(trial)) / max(trial) # normalized trial indices
  ) %>%
  rowwise() %>%
  mutate(interact = !any(is.na(unlist(interactions_processed)))) %>% # boolean to code f
or any interaction whatsoever
  unite("vis_interact", vis, interact, remove = FALSE)
```

Let’s exclude workers who miss the trial where causal support is the largest. We define miss as absolute error greater than 50%. This mean conditioning on only one of our attention checks to exclude about 22% of participants, rather than conditioning on both attention checks as preregistered which would exclude 48% of participants. 48% is too much, and this reflects the fact that we underestimated the difficulty of our second attention check trial, where causal support is at a minimum. Here, we are departing from our preregistration, but we are doing so in a way that admits more noise into our sample and is thus a more conservative analysis decision than the exclusion criteria we preregistered.

```
exclude_df <- model_df %>%
  group_by(workerId) %>%
  summarise(
    max_trial_idx = which(trial_idx == -1)[1],
    max_trial_gt = ground_truth[[max_trial_idx]],
    max_trial_err = abs_err[[max_trial_idx]],
    exclude = max_trial_err > 0.5
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(exclude_df)
```

```
## # A tibble: 6 x 5
##   workerId max_trial_idx max_trial_gt max_trial_err exclude
##   <chr>            <int>        <dbl>         <dbl> <lgl>
## 1 0030b402            13            1          0.3  FALSE
## 2 006327b8             7            1          0.3  FALSE
## 3 0116d604             7            1          0.4  FALSE
## 4 0306e6e7             7            1          0.2  FALSE
## 5 039a9f69             7            1          0.1  FALSE
## 6 03fd4b52             7            1          0.55 TRUE
```

Apply the exclusion criteria.

```
model_df = exclude_df %>%
  select(workerId, exclude) %>%
  full_join(model_df, by = "workerId") %>%
  filter(!exclude) %>%
  select(-exclude)
```

Additionally, we'll drop all attention check trials now that we are done using them for exclusions. Because of a bug that inserted extra attention check trials for the first 135 workers, this means dropping more trials than it should for this subset of workers. Thus, we have more trials per participant after the bug was fixed.

```
model_df = model_df %>%
  filter(trial_idx != -1 & trial_idx != -2)
```

How many participants per condition after exclusions? (target sample size was 80 per condition)

```
model_df %>%
  group_by(vis) %>%
  summarise(
    n = length(unique(workerId))
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   vis          n
##   <fct>    <int>
## 1 aggbars     81
## 2 bars        86
## 3 filtbars    81
## 4 icons       81
## 5 text        80
```

We overshot our target sample size slightly in all but one condition. This happened because we launch HITs in batches on MTurk, and it is hard to anticipate how many people in a batch will pass the exclusion criterion. The few extra participants should not make much of a difference in our results.

Inferential model

This is the model that we will use statistical inferences, and it is the result of our preregistered model expansion process. See ModelExpansion.Rmd for more information about how we arrived at this model.

```
m <- brm(data = model_df, family = "gaussian",
  formula = bf(lrr ~ causal_support*delta_p*n*vis_interact + (causal_support*delta_p + causal_support*n|workerId)),
  prior = c(prior(normal(-0.1654036, 1), class = Intercept),         # center at mean(qlogis(model_df$response_A / 100))
            prior(normal(0, 0.5), class = b),                        # center predictor effects at 0
            prior(normal(1, 0.5), class = b, coef = causal_support), # center at unbiased slope
            prior(normal(0, 0.5), class = sigma),                    # weakly informative half-normal
            prior(normal(0, 0.5), class = sd),                       # weakly informative half-normal
            prior(lkj(4), class = cor)),                             # avoiding large correlations
  iter = 3000, warmup = 500, chains = 2, cores = 2,
  control = list(adapt_delta = 0.99, max_treedepth = 12),
  file = "model-fits/8_re-within")
```

```
summary(m)
```

```
##  Family: gaussian 
##   Links: mu = identity; sigma = identity 
## Formula: lrr ~ causal_support * delta_p * n * vis_interact + (causal_support * delta_
p + causal_support * n | workerId) 
##    Data: model_df (Number of observations: 6528) 
## Samples: 2 chains, each with iter = 3000; warmup = 500; thin = 1;
##          total post-warmup samples = 5000
## 
## Group-Level Effects: 
## ~workerId (Number of levels: 409) 
##                                                   Estimate Est.Error l-95% CI
## sd(Intercept)                                         0.88      0.04     0.80
## sd(causal_support)                                    0.11      0.01     0.09
## sd(delta_p)                                           4.37      0.27     3.83
## sd(n500)                                              0.23      0.06     0.10
## sd(n1000)                                             0.40      0.07     0.25
## sd(n1500)                                             0.44      0.07     0.29
## sd(causal_support:delta_p)                            0.53      0.06     0.41
## sd(causal_support:n500)                               0.02      0.01     0.00
## sd(causal_support:n1000)                              0.01      0.01     0.00
## sd(causal_support:n1500)                              0.03      0.01     0.01
## cor(Intercept,causal_support)                        -0.45      0.09    -0.63
## cor(Intercept,delta_p)                               -0.66      0.05    -0.74
## cor(causal_support,delta_p)                           0.74      0.08     0.58
## cor(Intercept,n500)                                   0.34      0.15     0.04
## cor(causal_support,n500)                             -0.37      0.15    -0.65
## cor(delta_p,n500)                                    -0.27      0.15    -0.56
## cor(Intercept,n1000)                                  0.20      0.13    -0.03
## cor(causal_support,n1000)                            -0.52      0.11    -0.72
## cor(delta_p,n1000)                                   -0.31      0.12    -0.56
## cor(n500,n1000)                                       0.57      0.16     0.18
## cor(Intercept,n1500)                                 -0.00      0.11    -0.21
## cor(causal_support,n1500)                            -0.56      0.10    -0.75
## cor(delta_p,n1500)                                   -0.32      0.12    -0.56
## cor(n500,n1500)                                       0.52      0.17     0.12
## cor(n1000,n1500)                                      0.67      0.11     0.41
## cor(Intercept,causal_support:delta_p)                 0.33      0.10     0.12
## cor(causal_support,causal_support:delta_p)           -0.92      0.03    -0.96
## cor(delta_p,causal_support:delta_p)                  -0.73      0.08    -0.87
## cor(n500,causal_support:delta_p)                      0.20      0.17    -0.15
## cor(n1000,causal_support:delta_p)                     0.36      0.14     0.08
## cor(n1500,causal_support:delta_p)                     0.44      0.13     0.18
## cor(Intercept,causal_support:n500)                   -0.25      0.21    -0.62
## cor(causal_support,causal_support:n500)               0.24      0.21    -0.22
## cor(delta_p,causal_support:n500)                      0.23      0.21    -0.22
## cor(n500,causal_support:n500)                        -0.27      0.24    -0.68
## cor(n1000,causal_support:n500)                       -0.17      0.23    -0.58
## cor(n1500,causal_support:n500)                       -0.20      0.23    -0.60
## cor(causal_support:delta_p,causal_support:n500)      -0.21      0.22    -0.60
## cor(Intercept,causal_support:n1000)                  -0.06      0.25    -0.53
## cor(causal_support,causal_support:n1000)             -0.03      0.24    -0.48
## cor(delta_p,causal_support:n1000)                     0.03      0.24    -0.44
## cor(n500,causal_support:n1000)                       -0.01      0.24    -0.48
## cor(n1000,causal_support:n1000)                     -0.10      0.25    -0.58
## cor(n1500,causal_support:n1000)                      0.05      0.24    -0.43
## cor(causal_support:delta_p,causal_support:n1000)     0.00      0.24    -0.46
## cor(causal_support:n500,causal_support:n1000)        0.06      0.25    -0.41
## cor(Intercept,causal_support:n1500)                  0.46      0.16     0.09
## cor(causal_support,causal_support:n1500)            -0.38      0.17    -0.66
## cor(delta_p,causal_support:n1500)                   -0.31      0.17    -0.61
## cor(n500,causal_support:n1500)                       0.26      0.21    -0.17
## cor(n1000,causal_support:n1500)                      0.35      0.19    -0.07
## cor(n1500,causal_support:n1500)                      0.00      0.20    -0.38
## cor(causal_support:delta_p,causal_support:n1500)     0.16      0.18    -0.21
## cor(causal_support:n500,causal_support:n1500)       -0.07      0.23    -0.52
## cor(causal_support:n1000,causal_support:n1500)       0.04      0.24    -0.43
##                                                   u-95% CI Rhat Bulk_ESS
## sd(Intercept)                                        0.96 1.00     2515
## sd(causal_support)                                   0.13 1.00     1360
## sd(delta_p)                                          4.89 1.00     1916
## sd(n500)                                             0.36 1.01      452
## sd(n1000)                                            0.53 1.00      435
## sd(n1500)                                            0.57 1.00      399
## sd(causal_support:delta_p)                           0.66 1.00     1446
## sd(causal_support:n500)                              0.05 1.00      872
## sd(causal_support:n1000)                             0.03 1.00      779
## sd(causal_support:n1500)                             0.04 1.00      790
## cor(Intercept,causal_support)                       -0.27 1.00     1472
## cor(Intercept,delta_p)                              -0.56 1.00     3474
## cor(causal_support,delta_p)                          0.88 1.01      935
## cor(Intercept,n500)                                  0.64 1.00     1607
## cor(causal_support,n500)                            -0.03 1.00     1288
## cor(delta_p,n500)                                    0.02 1.00     1672
## cor(Intercept,n1000)                                 0.46 1.00      836
## cor(causal_support,n1000)                           -0.28 1.00      905
## cor(delta_p,n1000)                                  -0.07 1.01      653
## cor(n500,n1000)                                      0.82 1.01      519
## cor(Intercept,n1500)                                 0.23 1.00     1138
## cor(causal_support,n1500)                           -0.34 1.00     1212
## cor(delta_p,n1500)                                  -0.09 1.00      756
## cor(n500,n1500)                                      0.78 1.01      498
## cor(n1000,n1500)                                     0.85 1.00      762
## cor(Intercept,causal_support:delta_p)                0.53 1.00     2559
## cor(causal_support,causal_support:delta_p)          -0.86 1.00     1974
## cor(delta_p,causal_support:delta_p)                 -0.56 1.01     1267
## cor(n500,causal_support:delta_p)                     0.53 1.00     1402
## cor(n1000,causal_support:delta_p)                    0.61 1.00     1294
## cor(n1500,causal_support:delta_p)                    0.68 1.00     1277
## cor(Intercept,causal_support:n500)                   0.21 1.00     3286
## cor(causal_support,causal_support:n500)              0.62 1.00     4369
## cor(delta_p,causal_support:n500)                     0.60 1.00     4048
## cor(n500,causal_support:n500)                        0.27 1.00     2693
## cor(n1000,causal_support:n500)                       0.29 1.00     3950
## cor(n1500,causal_support:n500)                       0.27 1.00     3874
## cor(causal_support:delta_p,causal_support:n500)      0.25 1.00     4285
## cor(Intercept,causal_support:n1000)                  0.43 1.00     3298
## cor(causal_support,causal_support:n1000)             0.44 1.00     5199
## cor(delta_p,causal_support:n1000)                    0.48 1.00     3852
## cor(n500,causal_support:n1000)                       0.46 1.00     4986
## cor(n1000,causal_support:n1000)                      0.40 1.00     2776
## cor(n1500,causal_support:n1000)                      0.52 1.00     6144
## cor(causal_support:delta_p,causal_support:n1000)     0.47 1.00     5209
## cor(causal_support:n500,causal_support:n1000)        0.53 1.00     2623
## cor(Intercept,causal_support:n1500)                  0.72 1.00     2368
## cor(causal_support,causal_support:n1500)            -0.00 1.00     2157
## cor(delta_p,causal_support:n1500)                    0.04 1.00     2035
## cor(n500,causal_support:n1500)                       0.62 1.00     1838
## cor(n1000,causal_support:n1500)                      0.67 1.00     1399
## cor(n1500,causal_support:n1500)                      0.40 1.00     2464
## cor(causal_support:delta_p,causal_support:n1500)     0.50 1.00     3746
## cor(causal_support:n500,causal_support:n1500)        0.38 1.00     3255
## cor(causal_support:n1000,causal_support:n1500)       0.50 1.00     3161
##                                                  Tail_ESS
## sd(Intercept)                                        3652
## sd(causal_support)                                   2578
## sd(delta_p)                                          3722
## sd(n500)                                              836
## sd(n1000)                                             712
## sd(n1500)                                             853
## sd(causal_support:delta_p)                           2653
## sd(causal_support:n500)                              1720
## sd(causal_support:n1000)                              852
## sd(causal_support:n1500)                              673
## cor(Intercept,causal_support)                        2688
## cor(Intercept,delta_p)                               3524
## cor(causal_support,delta_p)                          2136
## cor(Intercept,n500)                                  2580
## cor(causal_support,n500)                             2090
## cor(delta_p,n500)                                    3285
## cor(Intercept,n1000)                                 1682
## cor(causal_support,n1000)                            1602
## cor(delta_p,n1000)                                   1482
## cor(n500,n1000)                                       979
## cor(Intercept,n1500)                                 2052
## cor(causal_support,n1500)                            2121
## cor(delta_p,n1500)                                   2123
## cor(n500,n1500)                                       973
## cor(n1000,n1500)                                     1396
## cor(Intercept,causal_support:delta_p)                3407
## cor(causal_support,causal_support:delta_p)           2925
## cor(delta_p,causal_support:delta_p)                  2706
## cor(n500,causal_support:delta_p)                     2777
## cor(n1000,causal_support:delta_p)                    2394
## cor(n1500,causal_support:delta_p)                    2273
## cor(Intercept,causal_support:n500)                   3535
## cor(causal_support,causal_support:n500)              3707
## cor(delta_p,causal_support:n500)                     3880
## cor(n500,causal_support:n500)                        3671
## cor(n1000,causal_support:n500)                       3692
## cor(n1500,causal_support:n500)                       3610
## cor(causal_support:delta_p,causal_support:n500)      4099
## cor(Intercept,causal_support:n1000)                  3560
## cor(causal_support,causal_support:n1000)             4144
## cor(delta_p,causal_support:n1000)                    3689
## cor(n500,causal_support:n1000)                       3851
## cor(n1000,causal_support:n1000)                      3298
## cor(n1500,causal_support:n1000)                      3420
## cor(causal_support:delta_p,causal_support:n1000)     3871
## cor(causal_support:n500,causal_support:n1000)        3280
## cor(Intercept,causal_support:n1500)                  1374
## cor(causal_support,causal_support:n1500)             1950
## cor(delta_p,causal_support:n1500)                    2019
## cor(n500,causal_support:n1500)                       3133
## cor(n1000,causal_support:n1500)                      2209
## cor(n1500,causal_support:n1500)                      2971
## cor(causal_support:delta_p,causal_support:n1500)     4070
## cor(causal_support:n500,causal_support:n1500)        4162
## cor(causal_support:n1000,causal_support:n1500)       3253
## 
## Population-Level Effects:
##                                                         Estimate Est.Error
## Intercept                                                  -0.06      0.10
## causal_support                                              0.28      0.06
## delta_p                                                     2.02      0.39
## n500                                                       -0.33      0.10
## n1000                                                      -0.32      0.10
## n1500                                                      -0.37      0.11
## vis_interactaggbars_TRUE                                    0.01      0.15
## vis_interactbars_FALSE                                     -0.00      0.14
## vis_interactfiltbars_FALSE                                  0.24      0.16
## vis_interactfiltbars_TRUE                                   0.54      0.15
## vis_interacticons_FALSE                                    -0.46      0.14
## vis_interacttext_FALSE                                     -0.41      0.14
## causal_support:delta_p                                     -0.57      0.19
## causal_support:n500                                        -0.14      0.06
## causal_support:n1000                                       -0.18      0.06
## causal_support:n1500                                       -0.18      0.06
## delta_p:n500                                                0.30      0.46
## delta_p:n1000                                               0.40      0.47
## delta_p:n1500                                               0.43      0.47
## causal_support:vis_interactaggbars_TRUE                    -0.05      0.10
## causal_support:vis_interactbars_FALSE                       0.09      0.08
## causal_support:vis_interactfiltbars_FALSE                  -0.29      0.10
## causal_support:vis_interactfiltbars_TRUE                   -0.17      0.09
## causal_support:vis_interacticons_FALSE                      0.09      0.08
## causal_support:vis_interacttext_FALSE                      -0.09      0.08
## delta_p:vis_interactaggbars_TRUE                            0.08      0.50
## delta_p:vis_interactbars_FALSE                              0.57      0.47
## delta_p:vis_interactfiltbars_FALSE                         -0.29      0.47
## delta_p:vis_interactfiltbars_TRUE                           0.22      0.48
## delta_p:vis_interacticons_FALSE                             0.83      0.48
## delta_p:vis_interacttext_FALSE                              0.20      0.46
## n500:vis_interactaggbars_TRUE                              -0.15      0.18
## n1000:vis_interactaggbars_TRUE                             -0.03      0.19
## n1500:vis_interactaggbars_TRUE                             -0.05      0.19
## n500:vis_interactbars_FALSE                                 0.17      0.13
## n1000:vis_interactbars_FALSE                               -0.19      0.14
## n1500:vis_interactbars_FALSE                               -0.19      0.15
## n500:vis_interactfiltbars_FALSE                             0.47      0.17
## n1000:vis_interactfiltbars_FALSE                            0.38      0.18
## n1500:vis_interactfiltbars_FALSE                            0.46      0.19
## n500:vis_interactfiltbars_TRUE                              0.30      0.15
## n1000:vis_interactfiltbars_TRUE                             0.16      0.16
## n1500:vis_interactfiltbars_TRUE                             0.42      0.17
## n500:vis_interacticons_FALSE                                0.06      0.14
## n1000:vis_interacticons_FALSE                              -0.09      0.15
## n1500:vis_interacticons_FALSE                              -0.12      0.15
## n500:vis_interacttext_FALSE                                 0.23      0.13
## n1000:vis_interacttext_FALSE                                0.29      0.14
## n1500:vis_interacttext_FALSE                                0.42      0.15
## causal_support:delta_p:n500                                 0.03      0.22
## causal_support:delta_p:n1000                                0.21      0.21
## causal_support:delta_p:n1500                                0.12      0.20
## causal_support:delta_p:vis_interactaggbars_TRUE             0.22      0.28
## causal_support:delta_p:vis_interactbars_FALSE              -0.31      0.24
## causal_support:delta_p:vis_interactfiltbars_FALSE           0.46      0.28
## causal_support:delta_p:vis_interactfiltbars_TRUE            0.33      0.27
## causal_support:delta_p:vis_interacticons_FALSE             -0.33      0.25
## causal_support:delta_p:vis_interacttext_FALSE               0.04      0.25
## causal_support:n500:vis_interactaggbars_TRUE                0.03      0.10
## causal_support:n1000:vis_interactaggbars_TRUE              -0.03      0.10
## causal_support:n1500:vis_interactaggbars_TRUE               0.00      0.10
## causal_support:n500:vis_interactbars_FALSE                 -0.02      0.07
## causal_support:n1000:vis_interactbars_FALSE                -0.07      0.07
## causal_support:n1500:vis_interactbars_FALSE                -0.05      0.07
## causal_support:n500:vis_interactfiltbars_FALSE              0.11      0.09
## causal_support:n1000:vis_interactfiltbars_FALSE             0.19      0.09
## causal_support:n1500:vis_interactfiltbars_FALSE             0.18      0.09
## causal_support:n500:vis_interactfiltbars_TRUE               0.09      0.09
## causal_support:n1000:vis_interactfiltbars_TRUE              0.11      0.08
## causal_support:n1500:vis_interactfiltbars_TRUE              0.08      0.08
## causal_support:n500:vis_interacticons_FALSE                -0.06      0.08
## causal_support:n1000:vis_interacticons_FALSE               -0.07      0.08
## causal_support:n1500:vis_interacticons_FALSE               -0.10      0.08
## causal_support:n500:vis_interacttext_FALSE                  0.13      0.08
## causal_support:n1000:vis_interacttext_FALSE                 0.11      0.08
## causal_support:n1500:vis_interacttext_FALSE                 0.12      0.08
## delta_p:n500:vis_interactaggbars_TRUE                       0.01      0.50
## delta_p:n1000:vis_interactaggbars_TRUE                      0.04      0.51
## delta_p:n1500:vis_interactaggbars_TRUE                      0.00      0.49
## delta_p:n500:vis_interactbars_FALSE                         0.18      0.50
## delta_p:n1000:vis_interactbars_FALSE                        0.13      0.50
## delta_p:n1500:vis_interactbars_FALSE                        0.05      0.49
## delta_p:n500:vis_interactfiltbars_FALSE                    -0.00      0.50
## delta_p:n1000:vis_interactfiltbars_FALSE                    0.01      0.50
## delta_p:n1500:vis_interactfiltbars_FALSE                    0.00      0.49
## delta_p:n500:vis_interactfiltbars_TRUE                      0.11      0.49
## delta_p:n1000:vis_interactfiltbars_TRUE                     0.01      0.49
## delta_p:n1500:vis_interactfiltbars_TRUE                     0.09      0.50
## delta_p:n500:vis_interacticons_FALSE                         0.11      0.50
## delta_p:n1000:vis_interacticons_FALSE                        0.14      0.50
## delta_p:n1500:vis_interacticons_FALSE                        0.01      0.50
## delta_p:n500:vis_interacttext_FALSE                          0.02      0.49
## delta_p:n1000:vis_interacttext_FALSE                         0.13      0.49
## delta_p:n1500:vis_interacttext_FALSE                         0.18      0.51
## causal_support:delta_p:n500:vis_interactaggbars_TRUE         0.24      0.37
## causal_support:delta_p:n1000:vis_interactaggbars_TRUE        0.05      0.34
## causal_support:delta_p:n1500:vis_interactaggbars_TRUE        0.14      0.33
## causal_support:delta_p:n500:vis_interactbars_FALSE          -0.20      0.29
## causal_support:delta_p:n1000:vis_interactbars_FALSE          0.24      0.28
## causal_support:delta_p:n1500:vis_interactbars_FALSE          0.08      0.28
## causal_support:delta_p:n500:vis_interactfiltbars_FALSE       0.13      0.36
## causal_support:delta_p:n1000:vis_interactfiltbars_FALSE     -0.16      0.33
## causal_support:delta_p:n1500:vis_interactfiltbars_FALSE      0.10      0.32
## causal_support:delta_p:n500:vis_interactfiltbars_TRUE       -0.02      0.34
## causal_support:delta_p:n1000:vis_interactfiltbars_TRUE      -0.09      0.31
## causal_support:delta_p:n1500:vis_interactfiltbars_TRUE       0.08      0.30
## causal_support:delta_p:n500:vis_interacticons_FALSE          0.29      0.30
## causal_support:delta_p:n1000:vis_interacticons_FALSE         0.20      0.28
## causal_support:delta_p:n1500:vis_interacticons_FALSE         0.44      0.28
## causal_support:delta_p:n500:vis_interacttext_FALSE          -0.07      0.30
## causal_support:delta_p:n1000:vis_interacttext_FALSE         -0.13      0.29
## causal_support:delta_p:n1500:vis_interacttext_FALSE         -0.28      0.28
##                                                          l-95% CI u-95% CI Rhat
## Intercept                                                   -0.26     0.15 1.00
## causal_support                                               0.16     0.40 1.00
## delta_p                                                      1.25     2.79 1.00
## n500                                                        -0.53    -0.13 1.00
## n1000                                                       -0.53    -0.12 1.00
## n1500                                                       -0.59    -0.15 1.00
## vis_interactaggbars_TRUE                                    -0.30     0.31 1.00
## vis_interactbars_FALSE                                      -0.28     0.26 1.00
## vis_interactfiltbars_FALSE                                  -0.08     0.56 1.00
## vis_interactfiltbars_TRUE                                    0.23     0.84 1.00
## vis_interacticons_FALSE                                     -0.72    -0.18 1.00
## vis_interacttext_FALSE                                      -0.69    -0.14 1.00
## causal_support:delta_p                                      -0.93    -0.22 1.00
## causal_support:n500                                         -0.26    -0.03 1.00
## causal_support:n1000                                        -0.30    -0.07 1.00
## causal_support:n1500                                        -0.29    -0.07 1.00
## delta_p:n500                                                -0.62     1.23 1.00
## delta_p:n1000                                               -0.52     1.31 1.00
## delta_p:n1500                                               -0.52     1.34 1.00
## causal_support:vis_interactaggbars_TRUE                     -0.24     0.13 1.00
## causal_support:vis_interactbars_FALSE                       -0.07     0.24 1.00
## causal_support:vis_interactfiltbars_FALSE                   -0.48    -0.11 1.00
## causal_support:vis_interactfiltbars_TRUE                    -0.34     0.01 1.00
## causal_support:vis_interacticons_FALSE                      -0.07     0.26 1.00
## causal_support:vis_interacttext_FALSE                       -0.25     0.07 1.00
## delta_p:vis_interactaggbars_TRUE                            -0.89     1.05 1.00
## delta_p:vis_interactbars_FALSE                              -0.34     1.48 1.00
## delta_p:vis_interactfiltbars_FALSE                          -1.22     0.66 1.00
## delta_p:vis_interactfiltbars_TRUE                           -0.73     1.18 1.00
## delta_p:vis_interacticons_FALSE                         -0.10       1.75 1.00
## delta_p:vis_interacttext_FALSE                          -0.71       1.10 1.00
## n500:vis_interactaggbars_TRUE                           -0.49       0.20 1.00
## n1000:vis_interactaggbars_TRUE                          -0.40       0.34 1.00
## n1500:vis_interactaggbars_TRUE                          -0.42       0.32 1.00
## n500:vis_interactbars_FALSE                             -0.08       0.42 1.00
## n1000:vis_interactbars_FALSE                            -0.46       0.08 1.00
## n1500:vis_interactbars_FALSE                            -0.47       0.10 1.00
## n500:vis_interactfiltbars_FALSE                          0.15       0.80 1.00
## n1000:vis_interactfiltbars_FALSE                         0.04       0.74 1.00
## n1500:vis_interactfiltbars_FALSE                         0.09       0.82 1.00
## n500:vis_interactfiltbars_TRUE                           0.01       0.59 1.00
## n1000:vis_interactfiltbars_TRUE                         -0.15       0.47 1.00
## n1500:vis_interactfiltbars_TRUE                          0.09       0.74 1.00
## n500:vis_interacticons_FALSE                            -0.21       0.33 1.00
## n1000:vis_interacticons_FALSE                           -0.37       0.20 1.00
## n1500:vis_interacticons_FALSE                           -0.42       0.17 1.00
## n500:vis_interacttext_FALSE                             -0.05       0.49 1.00
## n1000:vis_interacttext_FALSE                             0.01       0.57 1.00
## n1500:vis_interacttext_FALSE                             0.13       0.70 1.00
## causal_support:delta_p:n500                             -0.39       0.45 1.00
## causal_support:delta_p:n1000                            -0.19       0.60 1.00
## causal_support:delta_p:n1500                            -0.27       0.49 1.00
## causal_support:delta_p:vis_interactaggbars_TRUE         -0.33       0.76 1.00
## causal_support:delta_p:vis_interactbars_FALSE           -0.78       0.17 1.00
## causal_support:delta_p:vis_interactfiltbars_FALSE       -0.10       1.01 1.00
## causal_support:delta_p:vis_interactfiltbars_TRUE        -0.17       0.85 1.00
## causal_support:delta_p:vis_interacticons_FALSE          -0.82       0.16 1.00
## causal_support:delta_p:vis_interacttext_FALSE           -0.44       0.52 1.00
## causal_support:n500:vis_interactaggbars_TRUE            -0.16       0.23 1.00
## causal_support:n1000:vis_interactaggbars_TRUE           -0.22       0.17 1.00
## causal_support:n1500:vis_interactaggbars_TRUE           -0.19       0.19 1.00
## causal_support:n500:vis_interactbars_FALSE              -0.16       0.13 1.00
## causal_support:n1000:vis_interactbars_FALSE             -0.21       0.08 1.00
## causal_support:n1500:vis_interactbars_FALSE             -0.19       0.10 1.00
## causal_support:n500:vis_interactfiltbars_FALSE          -0.08       0.29 1.00
## causal_support:n1000:vis_interactfiltbars_FALSE          0.02       0.38 1.00
## causal_support:n1500:vis_interactfiltbars_FALSE          0.00       0.36 1.00
## causal_support:n500:vis_interactfiltbars_TRUE           -0.08       0.26 1.00
## causal_support:n1000:vis_interactfiltbars_TRUE          -0.06       0.27 1.00
## causal_support:n1500:vis_interactfiltbars_TRUE          -0.09       0.25 1.00
## causal_support:n500:vis_interacticons_FALSE             -0.22       0.09 1.00
## causal_support:n1000:vis_interacticons_FALSE            -0.22       0.09 1.00
## causal_support:n1500:vis_interacticons_FALSE            -0.25       0.05 1.00
## causal_support:n500:vis_interacttext_FALSE              -0.02       0.29 1.00
## causal_support:n1000:vis_interacttext_FALSE             -0.04       0.26 1.00
## causal_support:n1500:vis_interacttext_FALSE             -0.03       0.27 1.00
## delta_p:n500:vis_interactaggbars_TRUE                   -0.94       0.97 1.00
## delta_p:n1000:vis_interactaggbars_TRUE                  -0.98       1.03 1.00
## delta_p:n1500:vis_interactaggbars_TRUE                  -0.95       0.96 1.00
## delta_p:n500:vis_interactbars_FALSE                     -0.80       1.17 1.00
## delta_p:n1000:vis_interactbars_FALSE                    -0.84       1.10 1.00
## delta_p:n1500:vis_interactbars_FALSE                    -0.91       1.02 1.00
## delta_p:n500:vis_interactfiltbars_FALSE                 -0.96       0.96 1.00
## delta_p:n1000:vis_interactfiltbars_FALSE                   -0.97     0.97 1.00
## delta_p:n1500:vis_interactfiltbars_FALSE                   -0.99     0.96 1.00
## delta_p:n500:vis_interactfiltbars_TRUE                     -0.84     1.05 1.00
## delta_p:n1000:vis_interactfiltbars_TRUE                    -0.97     0.98 1.00
## delta_p:n1500:vis_interactfiltbars_TRUE                    -0.88     1.04 1.00
## delta_p:n500:vis_interacticons_FALSE                       -0.89     1.11 1.00
## delta_p:n1000:vis_interacticons_FALSE                      -0.84     1.10 1.00
## delta_p:n1500:vis_interacticons_FALSE                      -0.97     0.99 1.00
## delta_p:n500:vis_interacttext_FALSE                        -0.92     0.97 1.00
## delta_p:n1000:vis_interacttext_FALSE                       -0.82     1.06 1.00
## delta_p:n1500:vis_interacttext_FALSE                       -0.80     1.19 1.00
## causal_support:delta_p:n500:vis_interactaggbars_TRUE       -0.48     0.99 1.00
## causal_support:delta_p:n1000:vis_interactaggbars_TRUE      -0.60     0.71 1.00
## causal_support:delta_p:n1500:vis_interactaggbars_TRUE      -0.49     0.77 1.00
## causal_support:delta_p:n500:vis_interactbars_FALSE         -0.78     0.36 1.00
## causal_support:delta_p:n1000:vis_interactbars_FALSE        -0.31     0.81 1.00
## causal_support:delta_p:n1500:vis_interactbars_FALSE        -0.47     0.61 1.00
## causal_support:delta_p:n500:vis_interactfiltbars_FALSE     -0.61     0.84 1.00
## causal_support:delta_p:n1000:vis_interactfiltbars_FALSE    -0.81     0.50 1.00
## causal_support:delta_p:n1500:vis_interactfiltbars_FALSE    -0.52     0.73 1.00
## causal_support:delta_p:n500:vis_interactfiltbars_TRUE      -0.70     0.64 1.00
## causal_support:delta_p:n1000:vis_interactfiltbars_TRUE     -0.71     0.53 1.00
## causal_support:delta_p:n1500:vis_interactfiltbars_TRUE     -0.52     0.66 1.00
## causal_support:delta_p:n500:vis_interacticons_FALSE        -0.29     0.88 1.00
## causal_support:delta_p:n1000:vis_interacticons_FALSE       -0.36     0.75 1.00
## causal_support:delta_p:n1500:vis_interacticons_FALSE       -0.11     0.98 1.00
## causal_support:delta_p:n500:vis_interacttext_FALSE         -0.66     0.52 1.00
## causal_support:delta_p:n1000:vis_interacttext_FALSE        -0.69     0.43 1.00
## causal_support:delta_p:n1500:vis_interacttext_FALSE        -0.82     0.26 1.00
##                                                         Bulk_ESS Tail_ESS
## Intercept                                                   1718     2390
## causal_support                                               847     1691
## delta_p                                                     4303     3609
## n500                                                        2376     3144
## n1000                                                       2086     2999
## n1500                                                       2221     3261
## vis_interactaggbars_TRUE                                    3206     3384
## vis_interactbars_FALSE                                      1903     2622
## vis_interactfiltbars_FALSE                                  2313     3278
## vis_interactfiltbars_TRUE                                   1914     3070
## vis_interacticons_FALSE                                     2079     3069
## vis_interacttext_FALSE                                      1917     2514
## causal_support:delta_p                                      2330     3284
## causal_support:n500                                         1129     2273
## causal_support:n1000                                         903     2087
## causal_support:n1500                                         864     1933
## delta_p:n500                                                8677     3627
## delta_p:n1000                                              11872     4162
## delta_p:n1500                                              12883     3784
## causal_support:vis_interactaggbars_TRUE                     2023     3055
## causal_support:vis_interactbars_FALSE                       1266     2233
## causal_support:vis_interactfiltbars_FALSE                   1370     2776
## causal_support:vis_interactfiltbars_TRUE                    1193     2740
## causal_support:vis_interacticons_FALSE                      1092     2028
## causal_support:vis_interacttext_FALSE                       1361     2355
## delta_p:vis_interactaggbars_TRUE                            9564     3934
## delta_p:vis_interactbars_FALSE                              7242     4018
## delta_p:vis_interactfiltbars_FALSE                          9365     3959
## delta_p:vis_interactfiltbars_TRUE                           7947     4077
## delta_p:vis_interacticons_FALSE                             7060     4128
## delta_p:vis_interacttext_FALSE                              7735     4184
## n500:vis_interactaggbars_TRUE                               4103     3661
## n1000:vis_interactaggbars_TRUE                              3624     3896
## n1500:vis_interactaggbars_TRUE                              3872     3925
## n500:vis_interactbars_FALSE                                 2899     3448
## n1000:vis_interactbars_FALSE                                2488     3536
## n1500:vis_interactbars_FALSE                                2666     3474
## n500:vis_interactfiltbars_FALSE                             4103     4368
## n1000:vis_interactfiltbars_FALSE                            3603     4244
## n1500:vis_interactfiltbars_FALSE                            3558     3952
## n500:vis_interactfiltbars_TRUE                              3374     3855
## n1000:vis_interactfiltbars_TRUE                             3028     3002
## n1500:vis_interactfiltbars_TRUE                             3174     3963
## n500:vis_interacticons_FALSE                                3016     3908
## n1000:vis_interacticons_FALSE                               2514     2978
## n1500:vis_interacticons_FALSE                               2653     3043
## n500:vis_interacttext_FALSE                                 3155     4143
## n1000:vis_interacttext_FALSE                                2606     3622
## n1500:vis_interacttext_FALSE                                2765     3512
## causal_support:delta_p:n500                                 3217     3799
## causal_support:delta_p:n1000                                3165     3784
## causal_support:delta_p:n1500                                3183     3974
## causal_support:delta_p:vis_interactaggbars_TRUE             4727     4070
## causal_support:delta_p:vis_interactbars_FALSE               2876     3192
## causal_support:delta_p:vis_interactfiltbars_FALSE           3985     3776
## causal_support:delta_p:vis_interactfiltbars_TRUE            3383     4066
## causal_support:delta_p:vis_interacticons_FALSE              2482     3814
## causal_support:delta_p:vis_interacttext_FALSE               2795     3438
## causal_support:n500:vis_interactaggbars_TRUE                2479     3718
## causal_support:n1000:vis_interactaggbars_TRUE               2169     2920
## causal_support:n1500:vis_interactaggbars_TRUE               2045     2962
## causal_support:n500:vis_interactbars_FALSE                  1602     3028
## causal_support:n1000:vis_interactbars_FALSE                 1493     2541
## causal_support:n1500:vis_interactbars_FALSE                 1363     2495
## causal_support:n500:vis_interactfiltbars_FALSE              1898     3110
## causal_support:n1000:vis_interactfiltbars_FALSE             1378     2750
## causal_support:n1500:vis_interactfiltbars_FALSE             1488     3067
## causal_support:n500:vis_interactfiltbars_TRUE               1623     2837
## causal_support:n1000:vis_interactfiltbars_TRUE              1634     2754
## causal_support:n1500:vis_interactfiltbars_TRUE              1304     2932
## causal_support:n500:vis_interacticons_FALSE                 1463     2716
## causal_support:n1000:vis_interacticons_FALSE                1196     2288
## causal_support:n1500:vis_interacticons_FALSE                1181     1961
## causal_support:n500:vis_interacttext_FALSE                  1760     2749
## causal_support:n1000:vis_interacttext_FALSE                 1541     2671
## causal_support:n1500:vis_interacttext_FALSE                 1359     2643
## delta_p:n500:vis_interactaggbars_TRUE                      11199     3489
## delta_p:n1000:vis_interactaggbars_TRUE                     13168     2691
## delta_p:n1500:vis_interactaggbars_TRUE                     12400     3345
## delta_p:n500:vis_interactbars_FALSE                        12633     3160
## delta_p:n1000:vis_interactbars_FALSE                        9809     3491
## delta_p:n1500:vis_interactbars_FALSE                       11978     3250
## delta_p:n500:vis_interactfiltbars_FALSE                    13117     3388
## delta_p:n1000:vis_interactfiltbars_FALSE                   12782     3334
## delta_p:n1500:vis_interactfiltbars_FALSE                   10705     3541
## delta_p:n500:vis_interactfiltbars_TRUE                      9111     3274
## delta_p:n1000:vis_interactfiltbars_TRUE                    12136     3140
## delta_p:n1500:vis_interactfiltbars_TRUE                    14013     3498
## delta_p:n500:vis_interacticons_FALSE                       10122     3137
## delta_p:n1000:vis_interacticons_FALSE                      12692     2918
## delta_p:n1500:vis_interacticons_FALSE                      11875     3249
## delta_p:n500:vis_interacttext_FALSE                        12998     3863
## delta_p:n1000:vis_interacttext_FALSE                       10415     3669
## delta_p:n1500:vis_interacttext_FALSE                       11650     3617
## causal_support:delta_p:n500:vis_interactaggbars_TRUE        8497     4071
## causal_support:delta_p:n1000:vis_interactaggbars_TRUE       6165     3882
## causal_support:delta_p:n1500:vis_interactaggbars_TRUE       5489     3782
## causal_support:delta_p:n500:vis_interactbars_FALSE          4235     4226
## causal_support:delta_p:n1000:vis_interactbars_FALSE         4131     4088
## causal_support:delta_p:n1500:vis_interactbars_FALSE         4012     3843
## causal_support:delta_p:n500:vis_interactfiltbars_FALSE      5980     4042
## causal_support:delta_p:n1000:vis_interactfiltbars_FALSE     5034     3890
## causal_support:delta_p:n1500:vis_interactfiltbars_FALSE     6142     4067
## causal_support:delta_p:n500:vis_interactfiltbars_TRUE       4990     3894
## causal_support:delta_p:n1000:vis_interactfiltbars_TRUE      5812     4097
## causal_support:delta_p:n1500:vis_interactfiltbars_TRUE      4854     4039
## causal_support:delta_p:n500:vis_interacticons_FALSE         4188     4046
## causal_support:delta_p:n1000:vis_interacticons_FALSE        4056     4143
## causal_support:delta_p:n1500:vis_interacticons_FALSE        4224     4471
## causal_support:delta_p:n500:vis_interacttext_FALSE          4018     4016
## causal_support:delta_p:n1000:vis_interacttext_FALSE         4126     3793
## causal_support:delta_p:n1500:vis_interacttext_FALSE         3882     4090
## 
## Family Specific Parameters: 
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.06      0.01     1.04     1.08 1.00     1546     3144
## 
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Main effects of visualization

Recall: We estimate the correspondence between user responses and our normative benchmark using a linear in log odds (LLO) model, where ideal performance is a one-to-one relationship between a user’s responses and normative causal support. We chartacterize performance primarily in terms of LLO slopes with respect to causal support, which is a measure of sensitivity to the signal in each data set that should support causal inferences. The LLO also has an intercept term, which measures the average response when there is no signal to support either causal explanation. Intercepts represent an overall bias in responses to the extent that they deviate from 50%.

To start, lets derive slopes and intercepts from our model.

```
# extract conditional expectations from model
results_df <- model_df %>%
  group_by(n, vis_interact, workerId) %>%
  data_grid(
    causal_support = c(0, 1),
    delta_p = quantile(model_df$delta_p, probs = plogis(seq(from = qlogis(0.001), to = q
logis(0.999), length.out = 20)))) %>%
  add_fitted_draws(m, value = "lrr_rep", seed = 1234, n = 500, re_formula = NA) %>%
  select(-one_of(c(".row",".chain",".iteration")))
```

```
## Adding missing grouping variables: `.row`
```

```
# derive slopes
slopes_df <- results_df %>%
  compare_levels(lrr_rep, by = causal_support) %>%
  rename(slope = lrr_rep)

# derive intercepts and merge dataframes
results_df <- results_df %>%
  filter(causal_support == 0) %>%
  rename(intercept = lrr_rep) %>%
  full_join(slopes_df, by = c("n", "vis_interact", "workerId", "delta_p", ".draw"))
```

Let’s also set the level order for our visualization conditions for plotting.

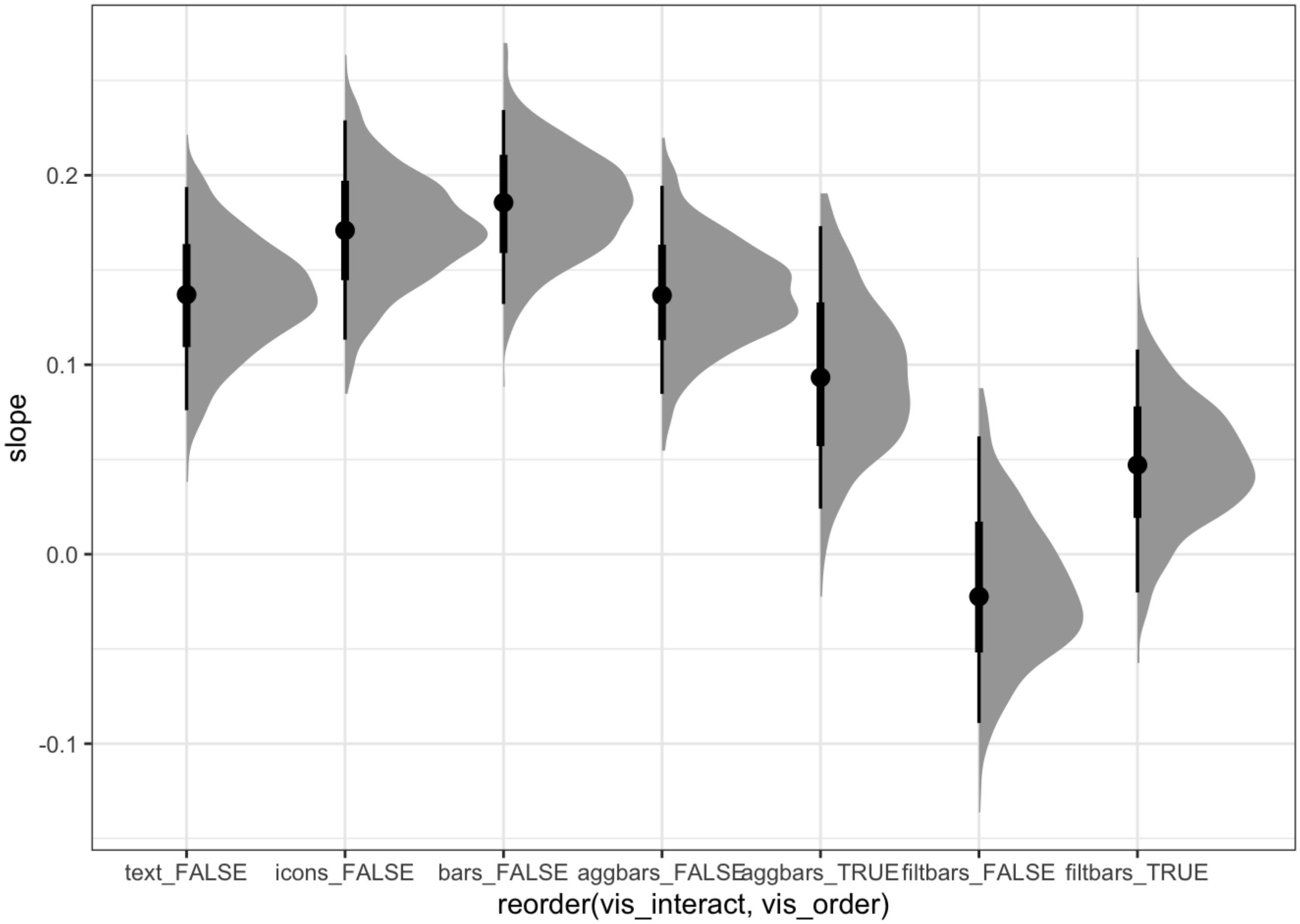
```
# relevel vis conditions to control plotting order
results_df <- results_df %>%
  mutate(
    vis_order = case_when(
      as.character(vis_interact) == "text_FALSE"     ~ 1,
      as.character(vis_interact) == "icons_FALSE"    ~ 2,
      as.character(vis_interact) == "bars_FALSE"     ~ 3,
      as.character(vis_interact) == "aggbars_FALSE"  ~ 4,
      as.character(vis_interact) == "aggbars_TRUE"   ~ 5,
      as.character(vis_interact) == "filtbars_FALSE" ~ 6,
      as.character(vis_interact) == "filtbars_TRUE"  ~ 7,
      TRUE                                           ~ 0
    ),
    vis_interact = reorder(vis_interact, vis_order)
  )
```

Let’s look at posterior estimates of the slope and intercept in each visualization condition.

We’ll start with *slopes*.

```
results_df %>%
  group_by(vis_interact, vis_order, .draw) %>% # group by predictors to keep
  summarise(slope = weighted.mean(slope)) %>%  # marginalize
  ggplot(aes(x = reorder(vis_interact, vis_order), y = slope)) +
    stat_halfeye() +
    theme_bw()
```

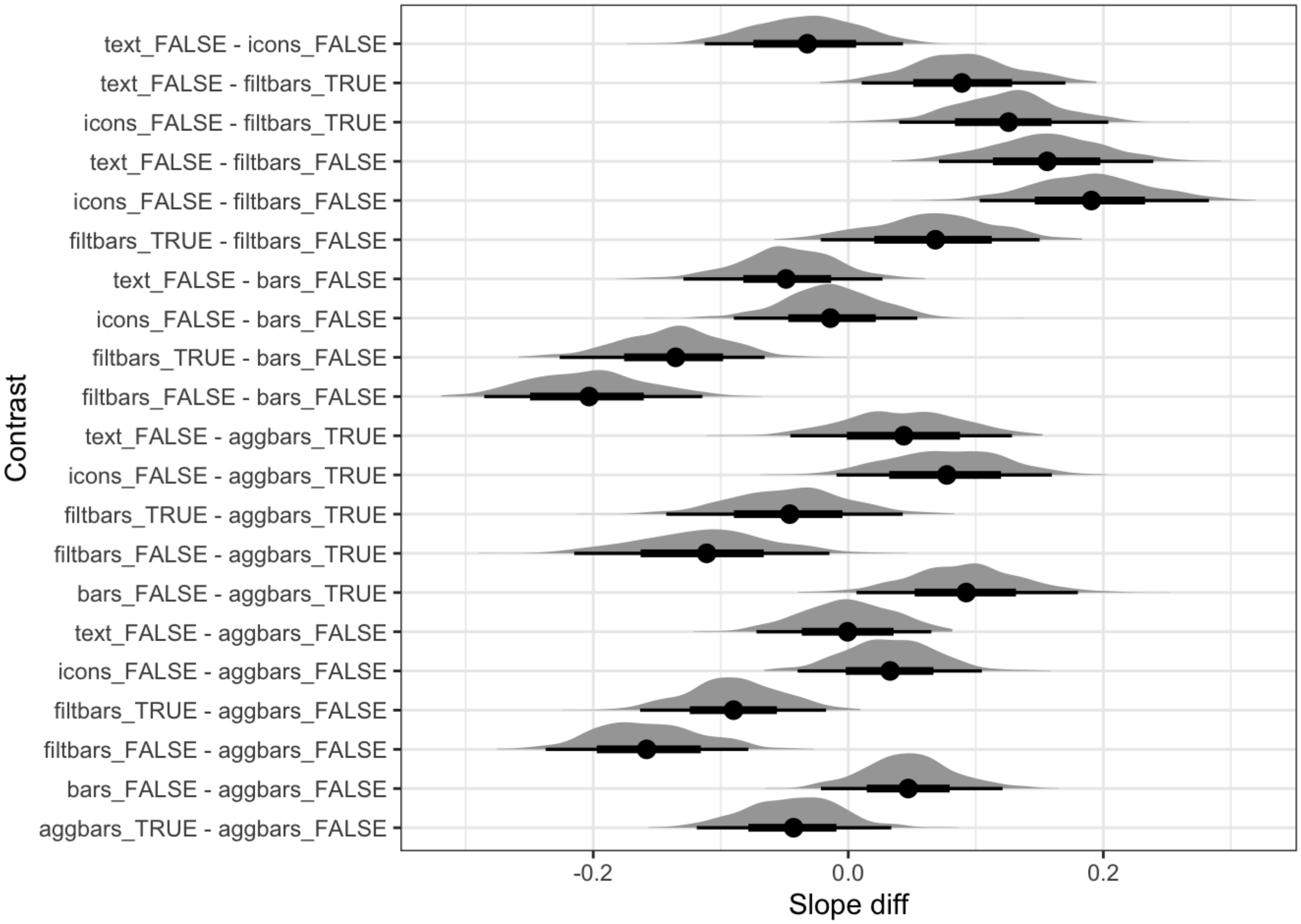
```
## `summarise()` regrouping output by 'vis_interact', 'vis_order' (override with `.group
s` argument)
```



Let's look at pairwise contrasts to see reliability of these visualization effects on LLO slopes. We'll flip the cooridinates so we have more space to put the labels for each pairwise difference.

```
slopes_df %>%
  group_by(vis_interact, .draw) %>%           # group by predictors to keep
  summarise(slope = weighted.mean(slope)) %>% # marginalize
  compare_levels(slope, by = vis_interact) %>%
  ggplot(aes(x = slope, y = vis_interact)) +
    stat_halfeyeh() +
    theme_bw() +
    labs(
      x = "Slope diff",
      y = "Contrast"
    )
```

```
## `summarise()` regrouping output by 'vis_interact' (override with `.groups` argument)
```



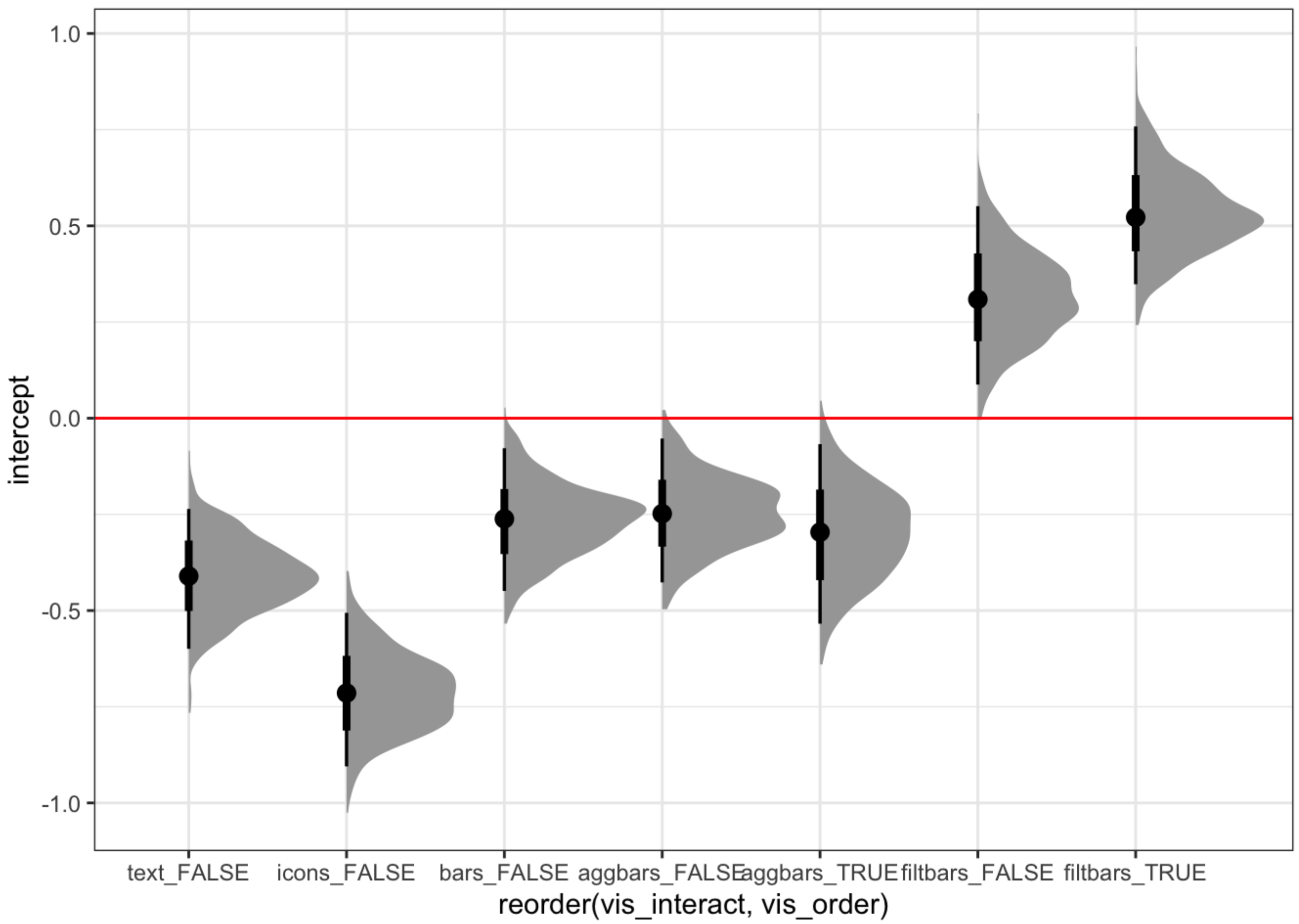
We can see that icons, bars, and text outperform the other visualization conditions with LLO slopes closer to 1. Differences between these conditions are not reliable. Overall, slopes are far from 1 in all conditions reflecting the difficulty of causal inference as a task.

Interestingly, bars without interaction reliably outperform aggbars and filtbars where people did interact. Users actually perform worse with aggbars when they do take the time to interact with the visualization which is unexpected, although this difference is not reliable. Users do pretty terrible with filtbars, but their performance improves when they take the time to interact, although this difference is also not reliable. Overall, chart users did reliably better with the table format visualizations than with the filtbars, which is expected considering that the information they need for the task is hidden behind clicks with filtbars.

Now, let’s look at *intercepts*.

```
results_df %>%
  group_by(vis_interact, vis_order, .draw) %>%        # group by predictors to keep
  summarise(intercept = weighted.mean(intercept)) %>% # marginalize
  ggplot(aes(x = reorder(vis_interact, vis_order), y = intercept)) +
    stat_halfeye() +
    geom_hline(yintercept = qlogis(0.5), color = "red") +
    theme_bw()
```

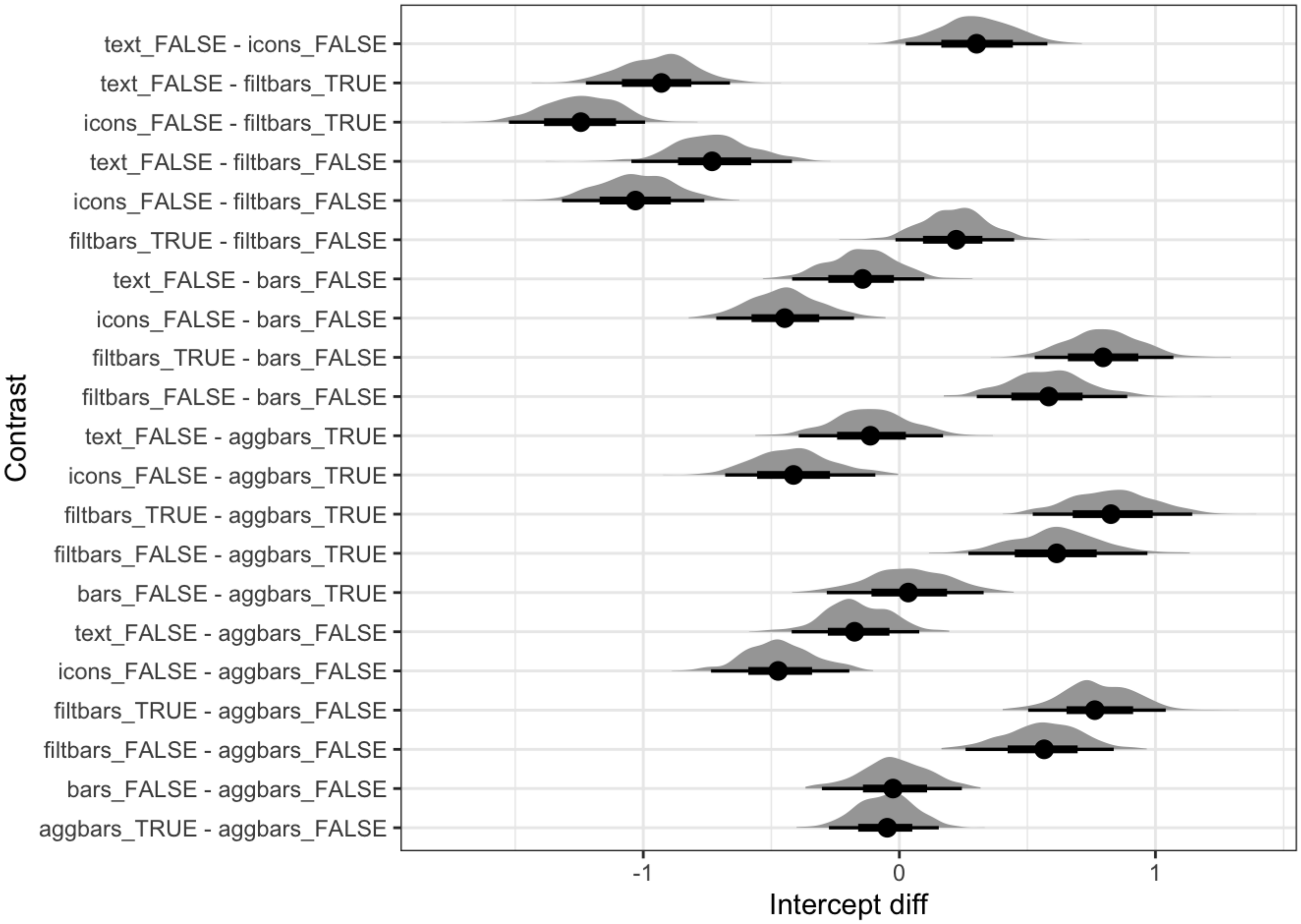
```
## `summarise()` regrouping output by 'vis_interact', 'vis_order' (override with `.group
s` argument)
```



As before, let’s look at pairwise contrasts to see reliability of these visualization effects on LLO intercepts.

```
results_df %>%
  group_by(vis_interact, .draw) %>%                    # group by predictors to keep
  summarise(intercept = weighted.mean(intercept)) %>% # marginalize
  compare_levels(intercept, by = vis_interact) %>%
  ggplot(aes(x = intercept, y = vis_interact)) +
    stat_halfeyeh() +
    theme_bw() +
    labs(
      x = "Intercept diff",
      y = "Contrast"
    )
```

```
## `summarise()` regrouping output by 'vis_interact' (override with `.groups` argument)
```



These intercepts show substantial response bias when there is no signal to support either causal explanation. People consistently overestimate the treatment effect with filtbars and underestimate it in all conditions using the table layout.

Icons lead to reliably more bias than text, bars, and aggbars. Text, bars, and aggbars are not reliably different from each other in term of bias.

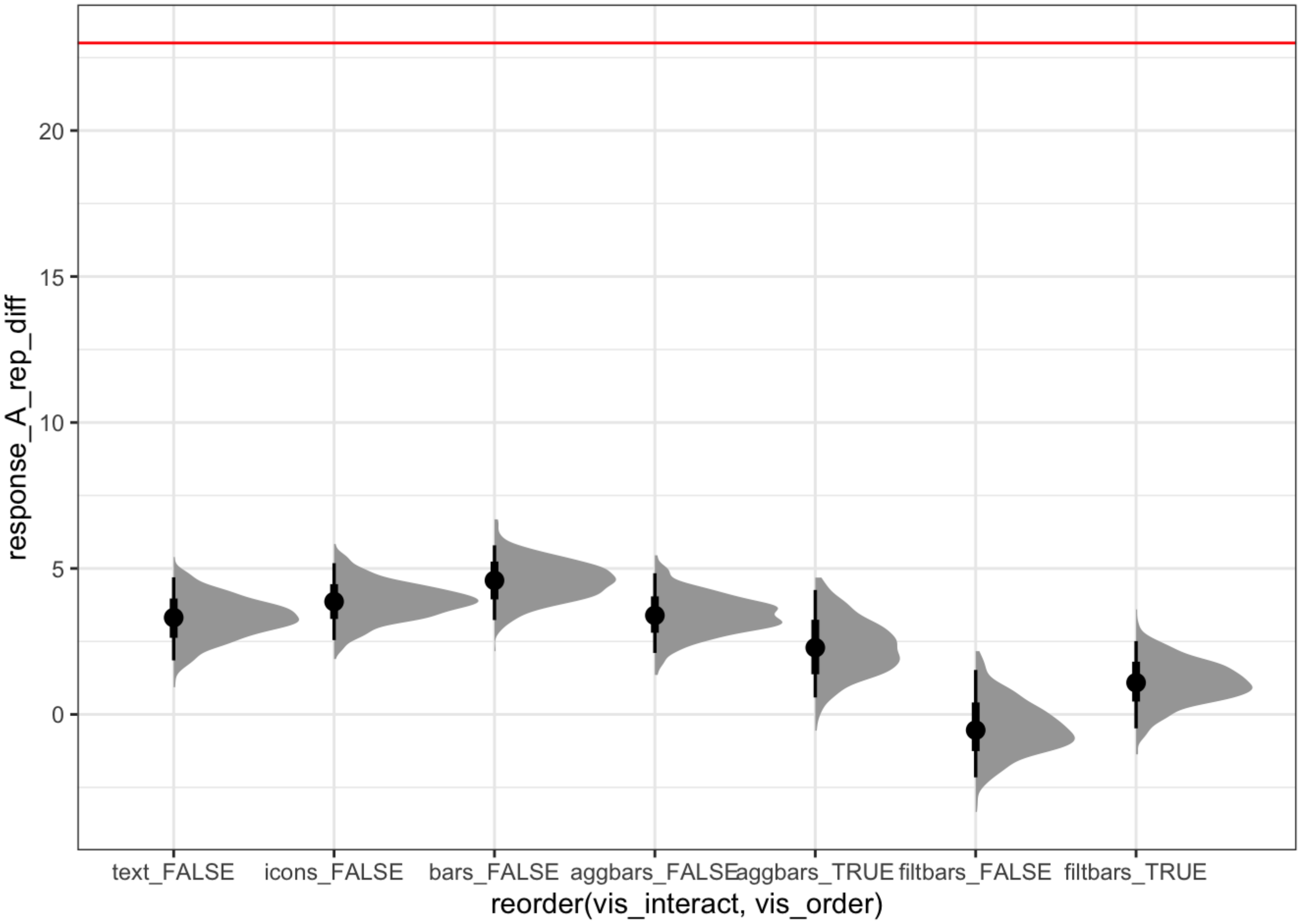
Effects of interaction are not reliable, but overestimation bias seems to increase when users interact with filtbars.

We can also frame these slopes and intercepts in terms of the response scale.

On the response scale, *slopes* are a change in the average user's subjective probability that there is a treatment effect given an increase in ground truth from `plogis(0) = 0.50` to `plogis(1) = 0.73`. A slope of 1 corresponds to an increase of 23% in the normative probability of a treatment effect.

```
results_df %>%
  group_by(vis_interact, vis_order, .draw) %>% # group by predictors to keep
  summarise(                                   # marginalize
    slope = weighted.mean(slope),
    intercept = weighted.mean(intercept)
  ) %>%
  mutate(
    response_A_rep_diff = (plogis(slope + intercept) - plogis(intercept)) * 100
  ) %>%
  ggplot(aes(x = reorder(vis_interact, vis_order), y = response_A_rep_diff)) +
    stat_halfeye() +
    geom_hline(yintercept = 23, color = "red") +
    theme_bw()
```

```
## `summarise()` regrouping output by 'vis_interact', 'vis_order' (override with `.group
s` argument)
```

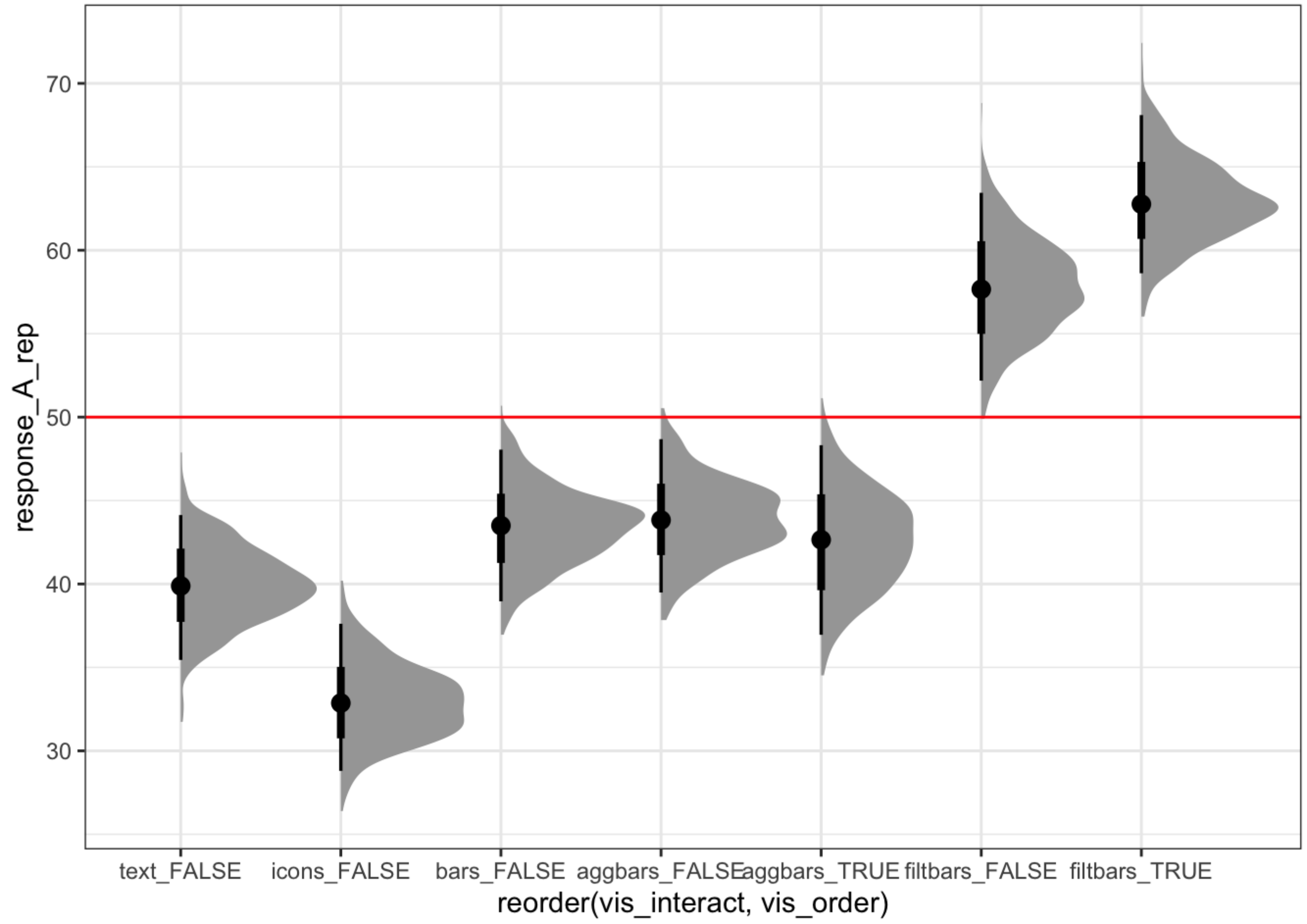


This view of the data reiterates that people are far less sensitive to the signal in the charts than they should be.

On the response scale, *intercepts* are just the average response where the ground truth is `plogis(0) = 0.5`. A response of 50% is ideal when there is no signal in the data.

```
results_df %>%
  group_by(vis_interact, vis_order, .draw) %>% # group by predictors to keep
  summarise(                                   # marginalize
    intercept = weighted.mean(intercept)
  ) %>%
  mutate(
    response_A_rep = plogis(intercept) * 100
  ) %>%
  ggplot(aes(x = reorder(vis_interact, vis_order), y = response_A_rep)) +
    stat_halfeye() +
    geom_hline(yintercept = 50, color = "red") +
    theme_bw()
```

```
## `summarise()` regrouping output by 'vis_interact', 'vis_order' (override with `.group
s` argument)
```



This view of the data helps us make sense of the magnitude of bias in the task. Users are over and underestimating the probability of a treatment effect by as much as 20% in some conditions. This is a very large amount of bias and was not necessarily something we expected to see (i.e., this was not even a preregistered comparison).

Interactions of visualization with delta p and sample size

In addition to how LLO slopes vary as a function of visualization condition, we want to investigate what aspects of the signal in a chart users seems to struggle to interpret. *The signal in our task for experiment 1 can be broken down into two attributes of the stimulus: delta p and sample size.*

Delta p is the difference in the proportion of people in each data set with the disease depending on whether they did vs didn't receive the treatment. Negative values of delta p indicate that a greater proportion of people had the disease in the treatment group than in the no treatment group (i.e., evidence against treatment effectiveness). Positive values of delta p indicate that a smaller proportion of people had the disease if they received treatment than if they didn't (i.e., evidence for treatment effectiveness).

Sample size is just the overall number of people in the fake data sets we showed on each trial.

In the ideal observer, there should be no residual effects of delta p and sample size after we've adjusted for the influence of causal support on user judgments. However, users have perceptual and cognitive biases in interpreting charts, which result in residual effects of delta p and sample size on user's responses.

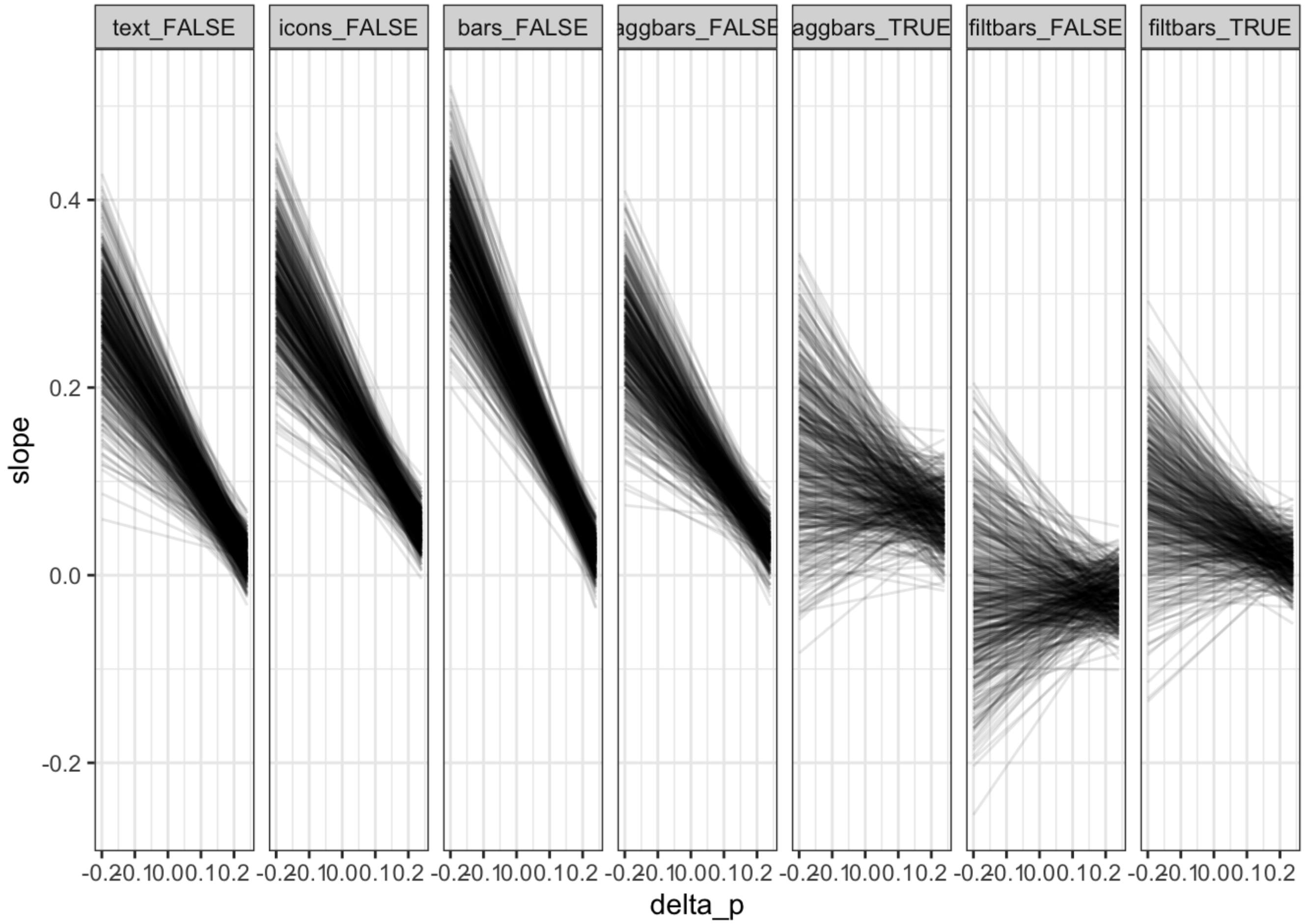
Here, we investigate preregistered comparisons of LLO slopes at different levels of delta p and sample size for each visualization condition. The degree to which LLO slopes deviate from one indicates how much these perceptual and cognitive biases distort sensitivity to the signal in charts.

Slopes

First, let's look at the *interaction between delta p and visualization condition on LLO slopes*. These lines should be flat with a y-intercept of 1 in an ideal observer.

```r
results_df %>%
  group_by(delta_p, vis_interact, vis_order, .draw) %>%  # group by predictors to keep
  summarise(slope = weighted.mean(slope)) %>%            # marginalize
  ggplot(aes(x = delta_p, y = slope, group = .draw)) +
    geom_line(alpha = 0.1) +
    theme_bw() +
    facet_grid(. ~ reorder(vis_interact, vis_order))
```

```
## `summarise()` regrouping output by 'delta_p', 'vis_interact', 'vis_order' (override w
ith `.groups` argument)
```

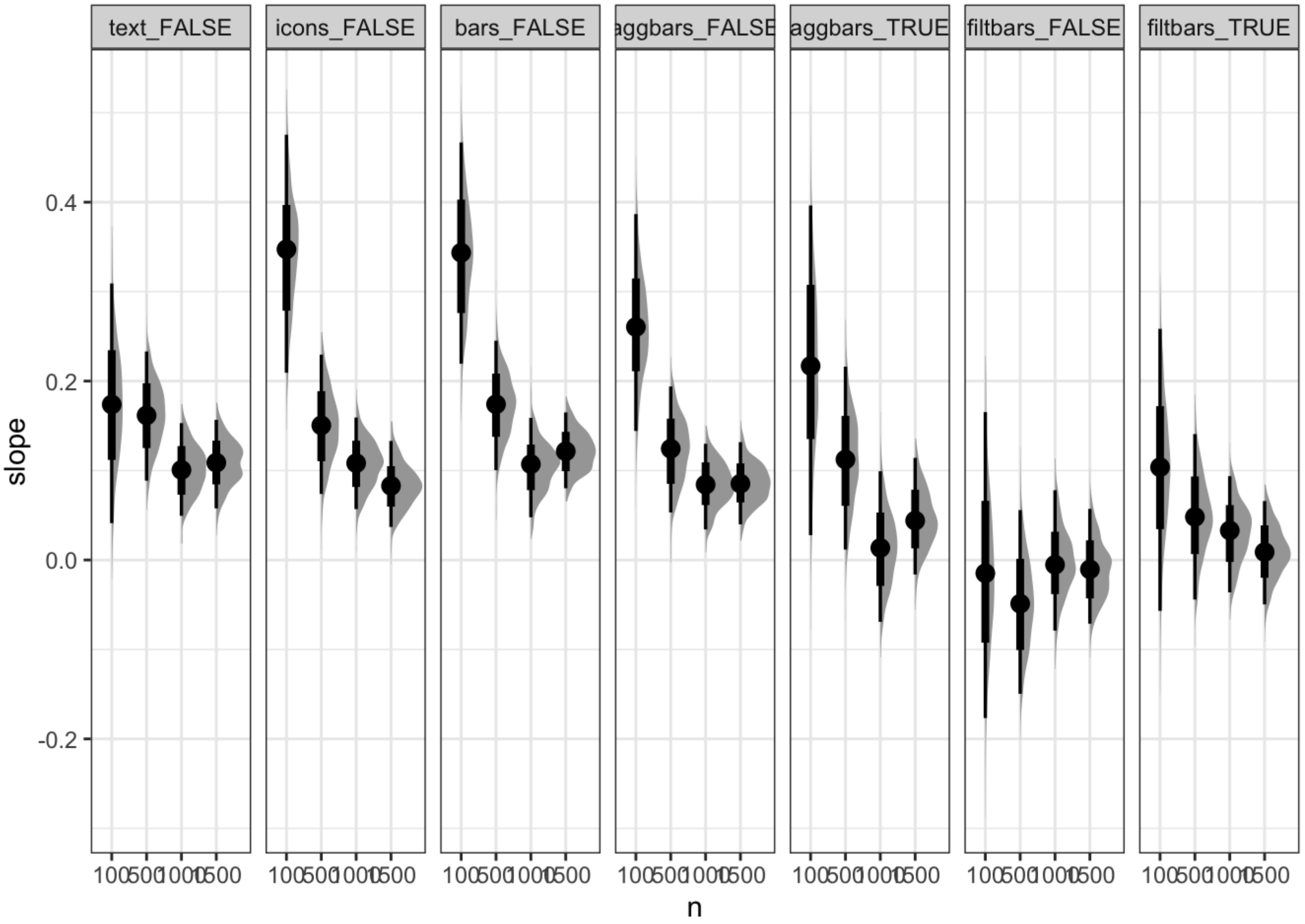


We can see that especially in the conditions with a tabular layout (i.e., aggbars, bars, icons, and text), users are more sensitive to signal (slopes closer to 1) in the charts when delta p is negative. This suggests that users are more sensitive to evidence against a treatment effect than evidence for one. Interestingly, this pattern seems to diminish when users interact with aggbars, with less sensitively at negative delta p, which may help to explain poorer performance when users interact with aggbars. Similarly, this pattern does not seem to happen much with filtbars, where sensitivity is much more uncertain at negative delta p.

Now, let’s look at the *interaction between sample size and visualization condition on LLO slopes*.

```
results_df %>%
  group_by(n, vis_interact, vis_order, .draw) %>% # group by predictors to keep
  summarise(slope = weighted.mean(slope)) %>%     # marginalize
  ggplot(aes(x = n, y = slope)) +
    stat_halfeye() +
    theme_bw() +
    facet_grid(. ~ reorder(vis_interact, vis_order))
```

```
## `summarise()` regrouping output by 'n', 'vis_interact', 'vis_order' (override with `.
groups` argument)
```



We see that users are less sensitive to the signal in charts as sample size increases, with the exception filtbars where performance is poor across the board. This trend is also less pronounced in the text condition. In particular, users seem to do best at low sample size, perhaps because the relatively small amount of data is a cue that strong inferences are not warranted. Although, this pattern may be more of a perceptual bias than a cognitive one. This result is consistent with prior work showing that people underestimate the number of items in a set and underestimate sample size for the purpose of making visual inferences with data.

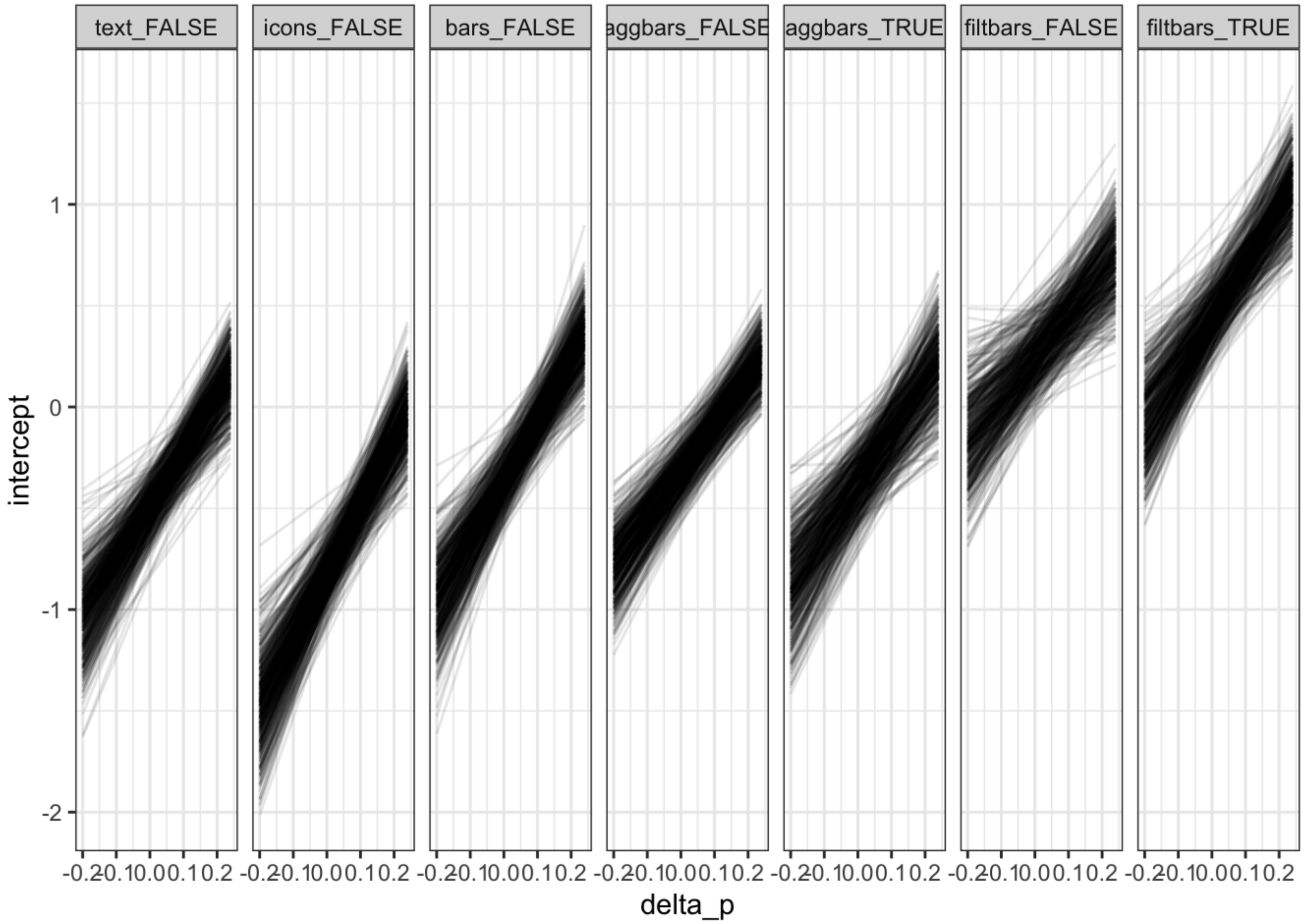
Intercepts

Although we did not preregister comparisons looking at the *interaction of visualization, delta p, and sample size on LLO intercepts*, the substantial amount of bias we saw in the intercept estimates per visualization condition make us curious.

We'll start by looking at the *interaction of delta p and visualization on LLO intercepts*. Astute readers will note that this is somewhat of a nonsensical counterfactual insofar as causal support depends in part on delta p, and extreme values of delta p seldom occur when ground truth causal support is 0 (i.e., at the intercept), with the exception of very very small sample sizes.

```
results_df %>%
  group_by(delta_p, vis_interact, vis_order, .draw) %>% # group by predictors to keep
  summarise(intercept = weighted.mean(intercept)) %>%   # marginalize
  ggplot(aes(x = delta_p, y = intercept, group = .draw)) +
    geom_line(alpha = 0.1) +
    theme_bw() +
    facet_grid(. ~ reorder(vis_interact, vis_order))
```

```
## `summarise()` regrouping output by 'delta_p', 'vis_interact', 'vis_order' (override w
ith `.groups` argument)
```

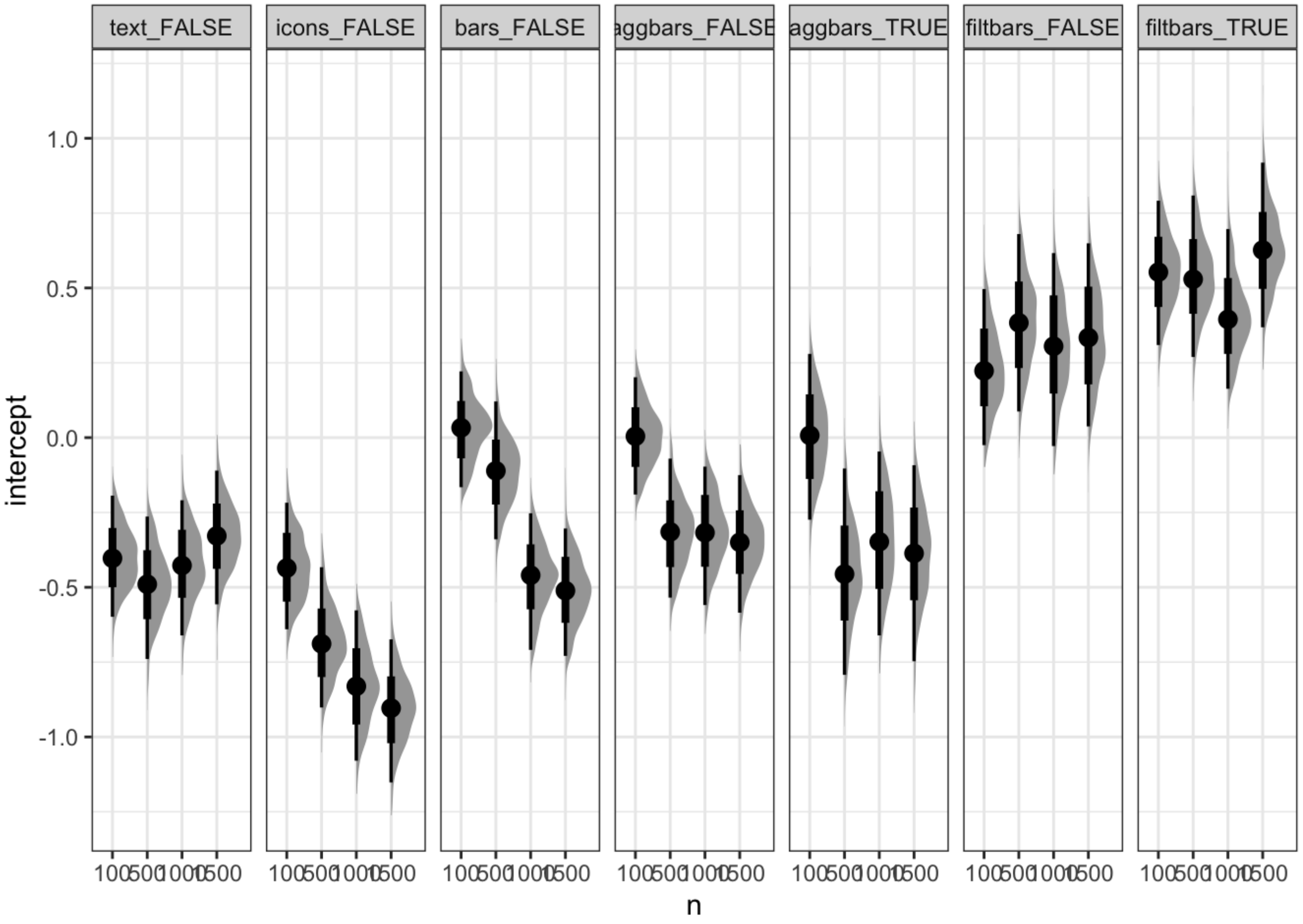


Here, we see what we expect to see if people understand the task. All else being equal, users say the treatment effect is more likely when the difference in the proportion of people with the disease suggests an effective treatment. This is a nice sanity check more than anything else.

Now, we'll look at the *interaction of sample size and visualization on LLO intercepts*. This query makes more sense than the last one since we can imagine scenarios at any sample size where the evidence for a treatment effect would appear totally ambiguous (i.e., ground truth causal support = 0).

```
results_df %>%
  group_by(n, vis_interact, vis_order, .draw) %>%     # group by predictors to keep
  summarise(intercept = weighted.mean(intercept)) %>% # marginalize
  ggplot(aes(x = n, y = intercept)) +
    stat_halfeye() +
    theme_bw() +
    facet_grid(. ~ reorder(vis_interact, vis_order))
```

```
## `summarise()` regrouping output by 'n', 'vis_interact', 'vis_order' (override with `.
groups` argument)
```



We can see that users tend to be the least biased in their responses at small sample sizes, especially with icons, bars, and aggbars. This bias to underestimate the probability of a treatment effect more at larger sample sizes is peculiar and unexpected but clearly a robust pattern.

User interactions with aggbars and filtbars

Let’s analyze how users interacted with the aggbars and filtbars visualization conditions, respectively.

We’ll start by writing functions to reconstruct the state of each visualization on each trial based on interaction logs.

```r
reconstruct_state_aggbars <- function(interactions) {
  # starting state is conditioning on both gene and treatment
  states <- list("gene_treat_init")

  for(i in 1:length(interactions)) {
    if (interactions[i] == "collapseRow" & str_detect(states[length(states)], "^gene_tre
at")) {
      states <- append(states, list("treat"))
    } else if (interactions[i] == "collapseRow" & states[length(states)] == "gene") {
      states <- append(states, list("none"))
    } else if (interactions[i] == "expandRow" & states[length(states)] == "treat") {
      states <- append(states, list("gene_treat"))
    } else if (interactions[i] == "expandRow" & states[length(states)] == "none") {
      states <- append(states, list("gene"))
    } else if (interactions[i] == "collapseCol" & str_detect(states[length(states)], "^g
ene_treat")) {
      states <- append(states, list("gene"))
    } else if (interactions[i] == "collapseCol" & states[length(states)] == "treat") {
      states <- append(states, list("none"))
    } else if (interactions[i] == "expandCol" & states[length(states)] == "gene") {
      states <- append(states, list("gene_treat"))
    } else if (interactions[i] == "expandCol" & states[length(states)] == "none") {
      states <- append(states, list("treat"))
    }
  }

  return(unlist(states))
}
```

```
reconstruct_state_filtbars <- function(interactions) {
  # starting state is conditioning on nothing
  states <- list("none_init")
  curr <- "" # state is a chain of filters

  for(i in 1:length(interactions)) {
    if (interactions[i] == "clearFilters") {
      curr <- ""
      states <- append(states, list("none"))
    } else if (str_detect(interactions[i], "^filter") & str_detect(states[length(state
s)], "^none")) {
      # first filter
      curr <- sub("^filter", "", interactions[i])
      states <- append(states, list(curr))
    } else if (str_detect(interactions[i], "^filter") & !str_detect(curr, paste(".*", su
b("^filter", "", interactions[i]), ".*", sep = ""))) {
      # only add interactions not already in the chain (don't log duplicate filters whic
h do not change the state)
      # put chain of filters including the current one into consistent order (so string
 matching can identify unique states)
      curr <- pmap_chr(list(curr, sub("^filter", "", interactions[i])), ~paste(sort(c(
...)), collapse = "_"))
      states <- append(states, list(curr))
    }
  }

  return(unlist(states))
}
```

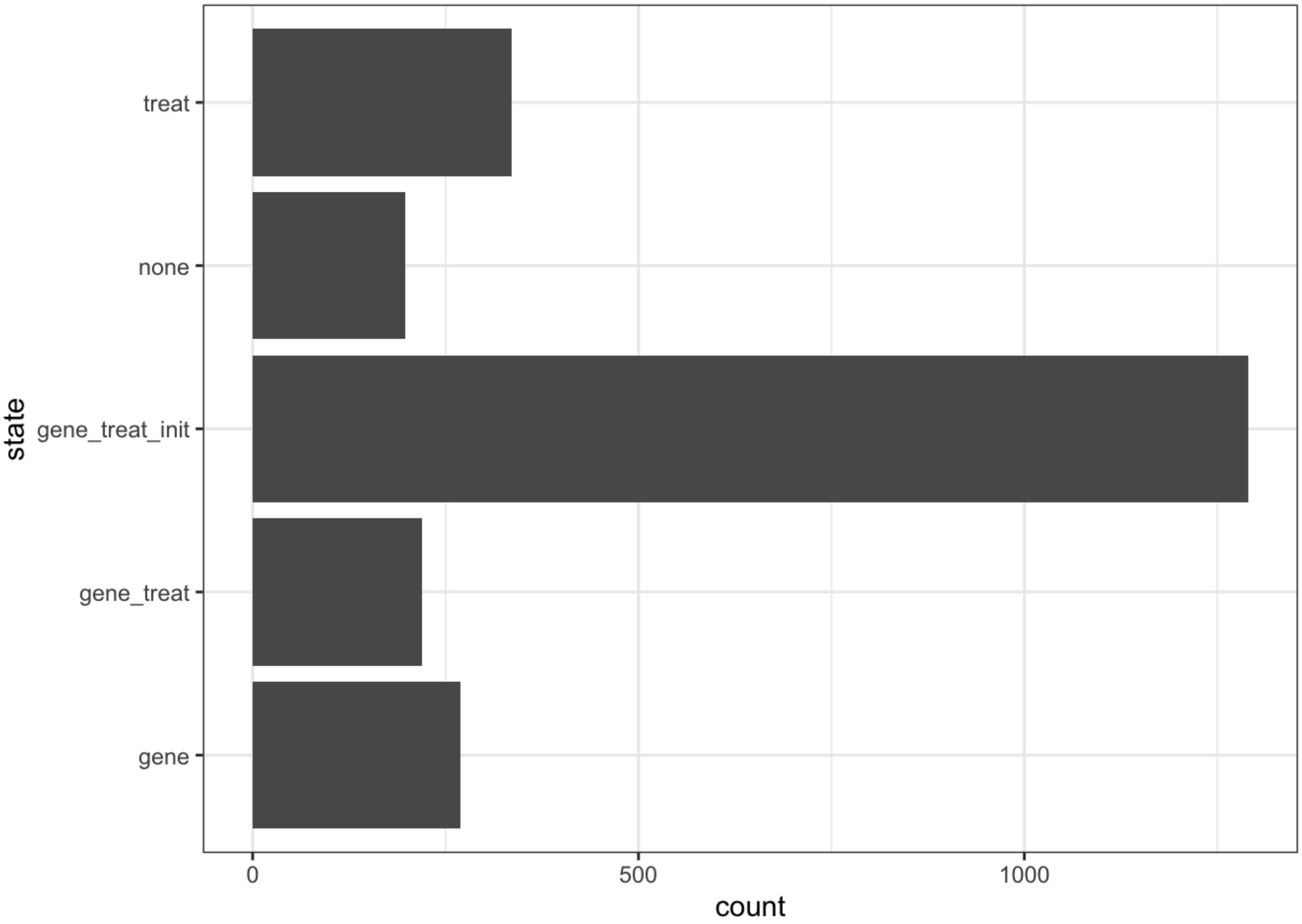
Now, we’ll reconstruct the states visited on each trial for each visualization separately.

```
aggbars_df <- model_df %>%
  filter(condition == "aggbars") %>%
  rowwise() %>%
  mutate(
    interactions = str_split(interactions, "_"),
    state = list(reconstruct_state_aggbars(interactions))
  )
```

```
filtbars_df <- model_df %>%
  filter(condition == "filtbars") %>%
  rowwise() %>%
  mutate(
    interactions = str_split(interactions, "_"),
    state = list(reconstruct_state_filtbars(interactions))
  )
```

Let’s view a histogram of the *states visited by users of aggbars*. These are named according to the conditions that users applied to the data in order to (dis)aggregate it.

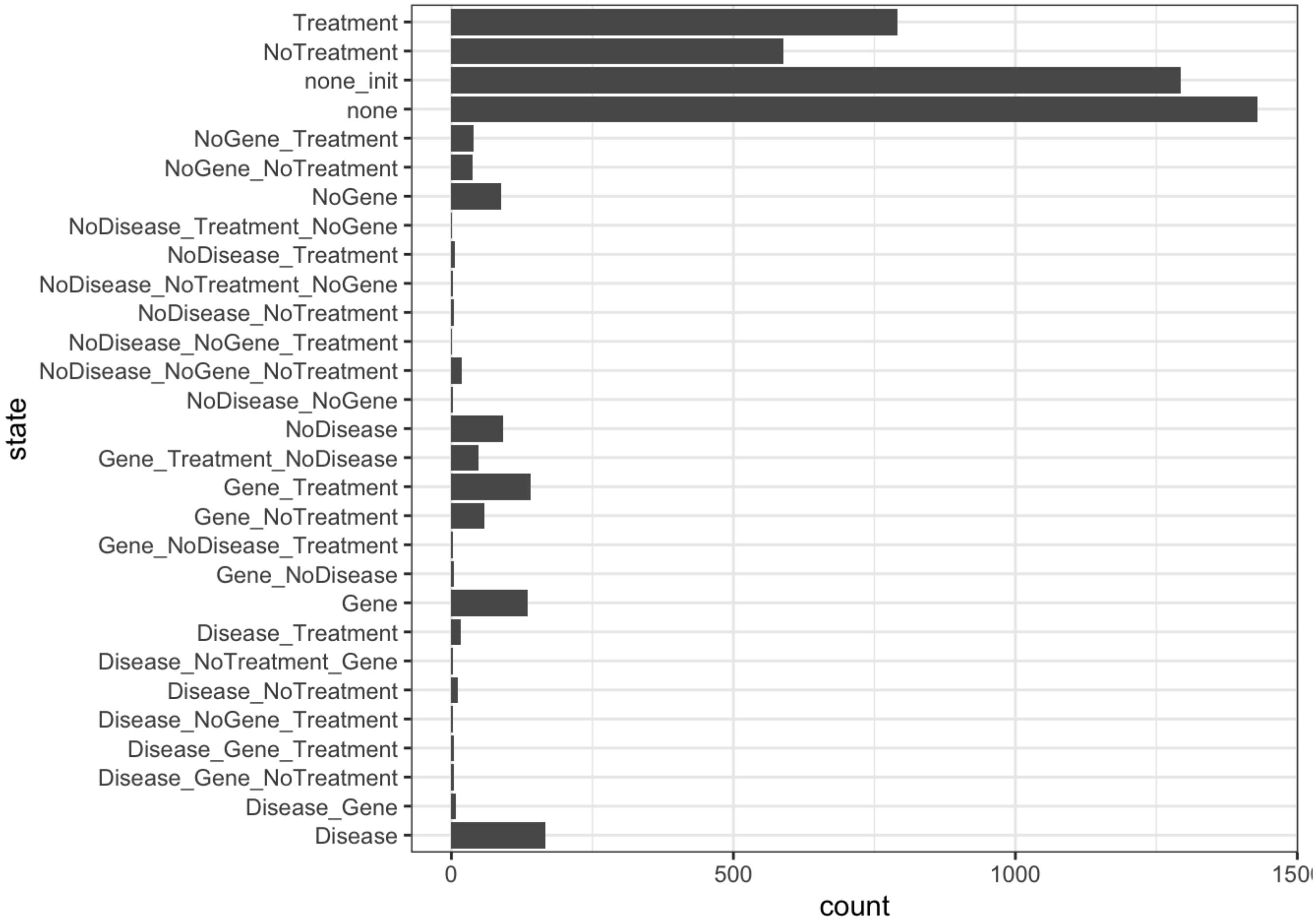
```
aggbars_df %>%
  group_by(trial, workerId) %>%
  unnest(cols = c("state")) %>%
  ggplot(aes(y = state)) +
  geom_bar() +
  theme_bw()
```



We can see that aggbars users create views conditioning on gene about as much as they create views conditioning on treatment.

Now, let’s view a histogram of the *states visited by users of filtbars*. These are named according to the conditions that users applied to the data in order to filter it.

```
filtbars_df %>%
  group_by(trial, workerId) %>%
  unnest(cols = c("state")) %>%
  ggplot(aes(y = state)) +
  geom_bar() +
  theme_bw()
```



We can see that users of filtbars are more likely to condition on treatment, if they interact with the visualization at all. Interestingly, some users also click the disease bar. While conditioning on the outcome variable is not a statistically valid one, it is a quick and intuitive way to see what other factors are most associated with getting the disease.

For both aggbars and filtbars, let’s see *what proportion of users create views that should be most helpful for the task*.

For aggbars this means intentionally creating views that condition on treatment, which we see in about 24% of trials.

```
aggbars_df <- aggbars_df %>%
  mutate(
    condition_on_treat = any(str_detect(unlist(state), ".*treat$"))
  )

sum(aggbars_df$condition_on_treat) / length(aggbars_df$condition_on_treat)
```

```
## [1] 0.2364341
```

For filtbars this means intentionally creating views that condition on both treatment and no treatment, which we see in about 33% of trials.

```
filtbars_df <- filtbars_df %>%
  mutate(
    condition_on_treat = any(str_detect(unlist(state), "^Treatment$")) & any(str_detect
(unlist(state), "^NoTreatment$"))
  )

sum(filtbars_df$condition_on_treat) / length(filtbars_df$condition_on_treat)
```

```
## [1] 0.3268934
```

The fact that people who interact with aggbars focus less on task relevant views than people who interact with filtbars may help to explain why interacting with aggbars is associated with worse performance while interacting with filtbars is associated with better performance.

Also, baseline performance with filtbars when people don’t interact is horrible, whereas baseline performance with aggbars when people don’t interact is decent. These different baselines may have something to do with the opposite directions of effect of interacting with the visualizations, since performance with filtbars could really only improve from baseline.