Data Corrections

A unique observation

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
results <- read_csv('../_data/stat218fall2025.csv')
```

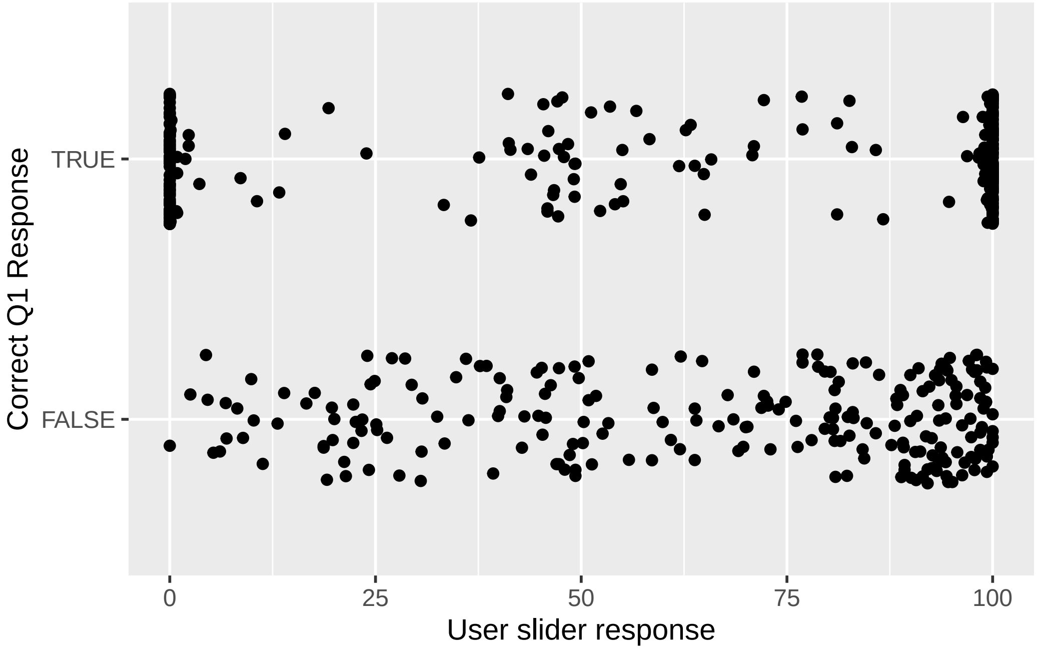
```
Rows: 4856 Columns: 31
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (11): set, media, user_id, user_larger, stage, user_age, user_gender, us...
dbl (17): pair_id, user_set_order, user_trial_order, user_slider, slider_sta...
lgl  (3): is_218_student, data_consent, q1_correct

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
solutions <- readRDS('../../data/solutions.rda')
```

An interesting result occurs when observing Stimuli Pair 5 – when participants correctly identified that the values were the same, three understandings of the task emerge.

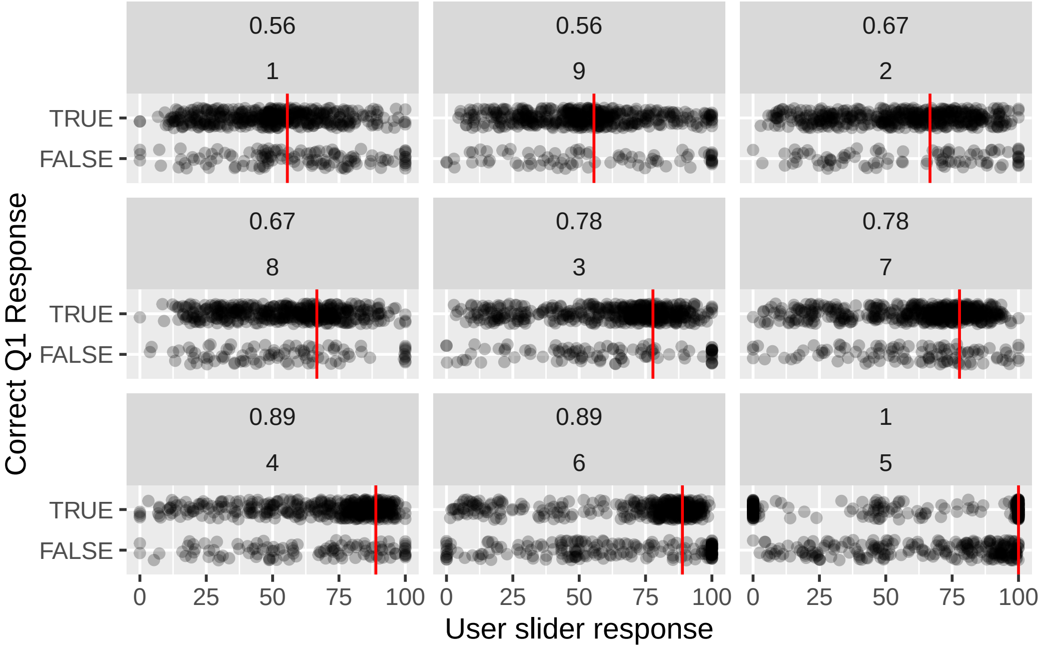
1. Task as assigned
2. True value estimation
3. Difference of stimuli

```
results %>%
  filter(pair_id == 5) %>%
  ggplot(mapping = aes(x = user_slider, y = q1_correct)) +
  geom_point(position = position_jitter(width = 0, height = 1/4)) +
  labs(x = 'User slider response', y = 'Correct Q1 Response')
```



This observation is not as obvious for the other stimuli pairs.

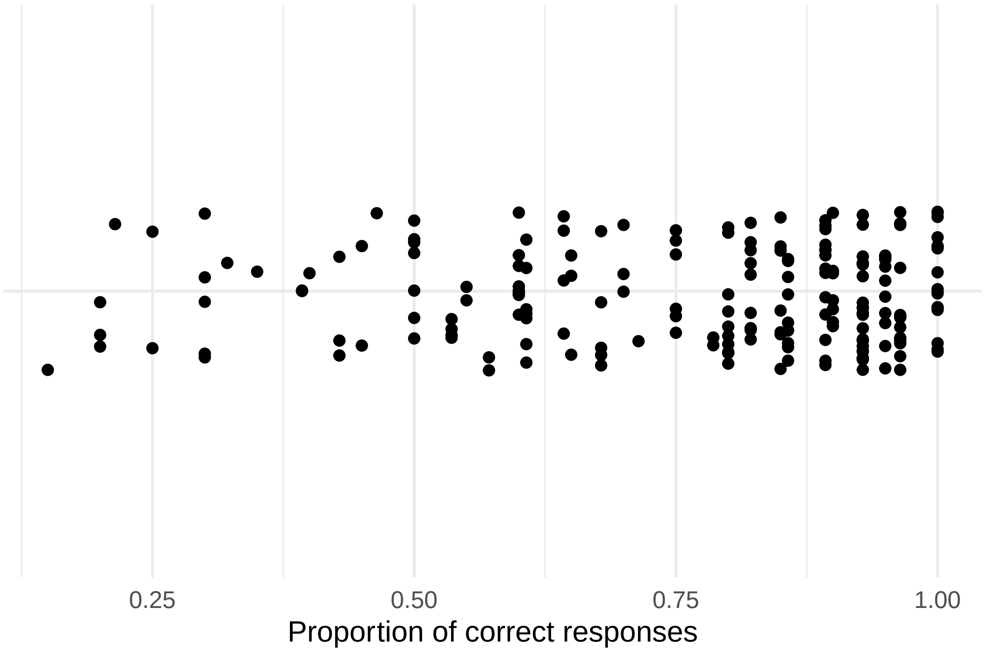
```
results %>%
  ggplot(mapping = aes(x = user_slider, y = q1_correct)) +
  geom_point(position = position_jitter(width = 0, height = 1/4),
             alpha = 1/4) +
  geom_vline(aes(xintercept = 100*true_ratio), color = 'red') +
  labs(x = 'User slider response', y = 'Correct Q1 Response') +
  facet_wrap(~round(true_ratio,2)+pair_id)
```



Participant tasks

If the task is consistent per participant, we should be able to correct the values. Looking at the correct proportion of correct responses for question 1 by participants, it appears that most participants were not nefarious is their submissions.

```
results %>%
  group_by(user_id) %>%
  summarize(prop_correct = mean(q1_correct)) %>%
  arrange(prop_correct) %>%
  ggplot(mapping = aes(x = prop_correct, y = '')) +
  geom_point(position = position_jitter(width = 0, height = 1/6)) +
  labs(x = 'Proportion of correct responses', y  = '') +
  theme_minimal()
```



For a consistent task nomenclature, refer to the following descriptions:

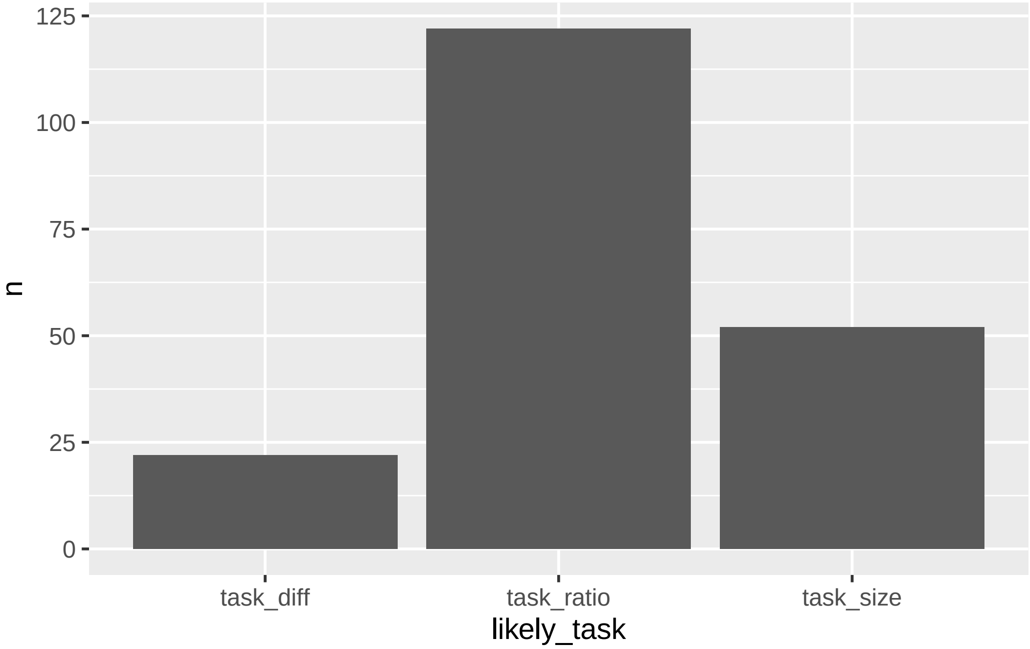
- Task Ratio: this is the actual task assigned, where participants estimate the size of the smaller value with respect to the larger value
- Task Size: this is an estimation of the value with respect to the chart
- Task Difference: this is where participants estimated the difference

A reasonable approach to figuring out which task was most likely to have occurred is to take the average difference between

```
load('../../data/stimuli.rda')
tasks <- results %>%
  left_join(stimuli) %>%
  mutate(smaller_value = map2_dbl(values, constant, min),
         larger_value = map2_dbl(values, constant, max),
         task_ratio = abs(user_slider - 100*true_ratio),
         task_size = abs(user_slider - smaller_value),
         task_diff = abs(user_slider - (larger_value - smaller_value)))
```

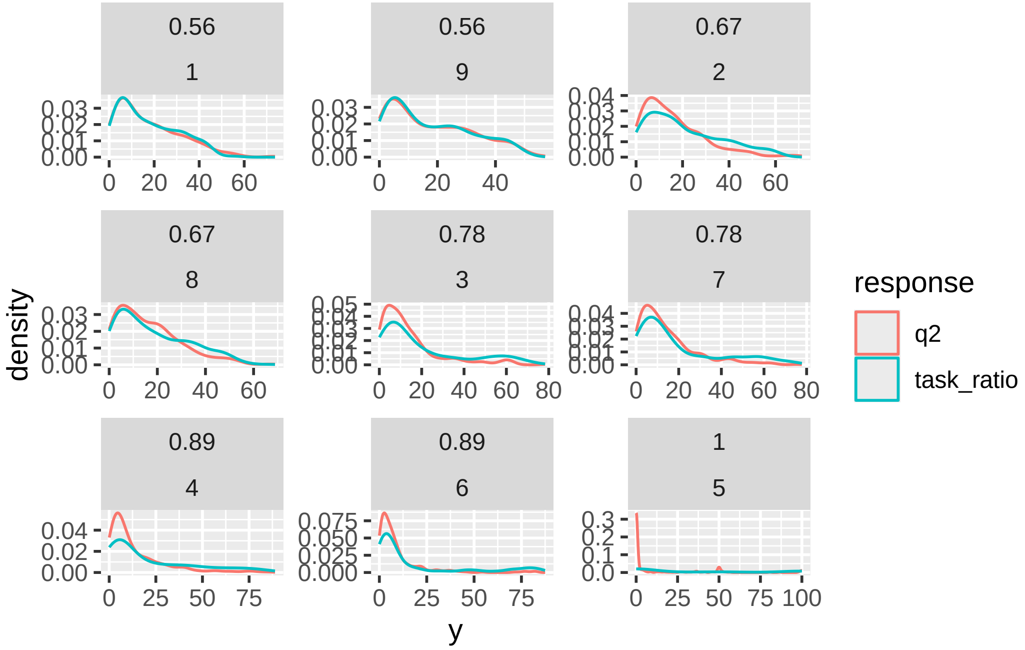
```
Joining with `by = join_by(pair_id)`
```

```
likely_task <- tasks %>%
  group_by(user_id) %>%
  filter(q1_correct) %>%
  summarise(task_ratio = mean(task_ratio),
            task_size = mean(task_size),
            task_diff = mean(task_diff)) %>%
  pivot_longer(task_ratio:task_diff, names_to = 'likely_task') %>%
  group_by(user_id) %>%
  filter(value == min(value))

likely_task %>%
  group_by(likely_task) %>%
  count() %>%
  ggplot(mapping = aes(x = likely_task, y = n)) +
  geom_col()
```



```
tasks %>%
  left_join(likely_task) %>%
  mutate(q2 = case_when(
    likely_task == 'task_diff' ~ task_diff,
    likely_task == 'task_size' ~ task_size,
    likely_task == 'task_ratio' ~ task_ratio
  )) %>%
  filter(q1_correct) %>%
  pivot_longer(c(task_ratio, q2), names_to = 'response', values_to = 'y') %>%
  ggplot(mapping = aes(x = y, color = response)) +
  geom_density(alpha = 1/2) +
  facet_wrap(~round(true_ratio,2)+pair_id, scales = 'free')
```

```
Joining with `by = join_by(user_id)`
```

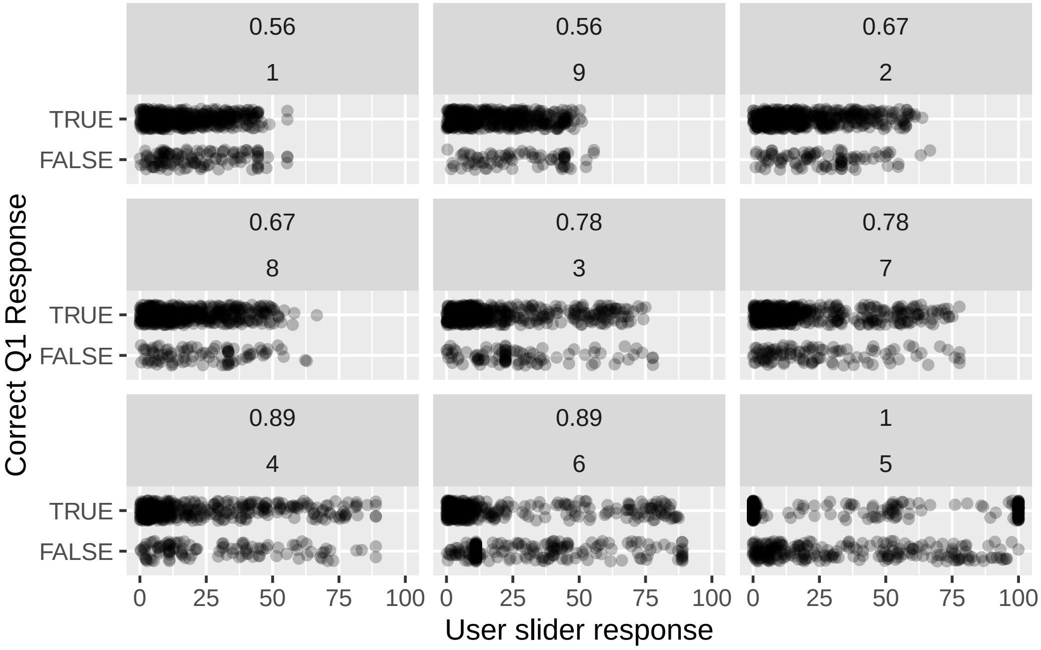


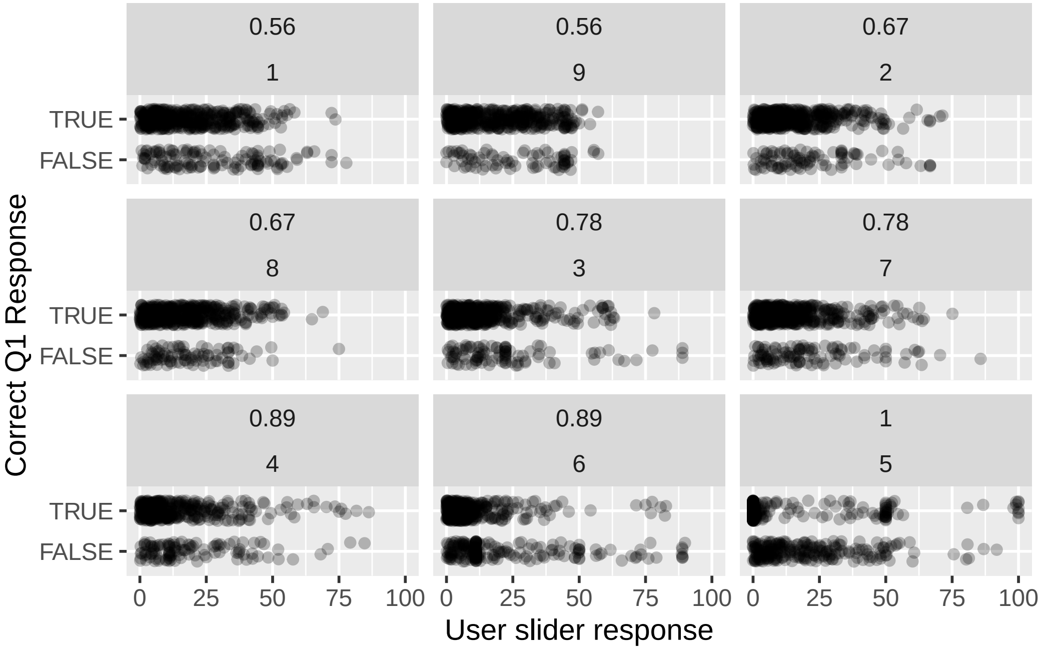
```
tasks %>%
  left_join(likely_task) %>%
  mutate(q2 = case_when(
    likely_task == 'task_diff' ~ task_diff,
    likely_task == 'task_size' ~ task_size,
    likely_task == 'task_ratio' ~ task_ratio
  ))  %>%
  ggplot(mapping = aes(x = q2_abs_error, y = q1_correct)) +
  geom_point(position = position_jitter(width = 0, height = 1/4),
             alpha = 1/4) +
  labs(x = 'User slider response', y = 'Correct Q1 Response') +
  facet_wrap(~round(true_ratio,2)+pair_id)
```

```
Joining with `by = join_by(user_id)`
```

```
tasks %>%
  left_join(likely_task) %>%
  mutate(q2 = case_when(
    likely_task == 'task_diff' ~ task_diff,
    likely_task == 'task_size' ~ task_size,
    likely_task == 'task_ratio' ~ task_ratio
  ))  %>%
  ggplot(mapping = aes(x = q2, y = q1_correct)) +
  geom_point(position = position_jitter(width = 0, height = 1/4),
             alpha = 1/4) +
  labs(x = 'User slider response', y = 'Correct Q1 Response') +
  facet_wrap(~round(true_ratio,2)+pair_id)
```

```
Joining with `by = join_by(user_id)`
```


0.56
1
0.56
9
0.67
2
0.67
8
0.78
3
0.78
7
0.89
4
0.89
6
1
5
TRUE
FALSE
Correct Q1 Response
0
25
50
75
100
User slider response


0.56
1
0.56
9
0.67
2
0.67
8
0.78
3
0.78
7
0.89
4
0.89
6
1
5
TRUE
FALSE
Correct Q1 Response
0
25
50
75
100
User slider response