Setup

Load necessary packages

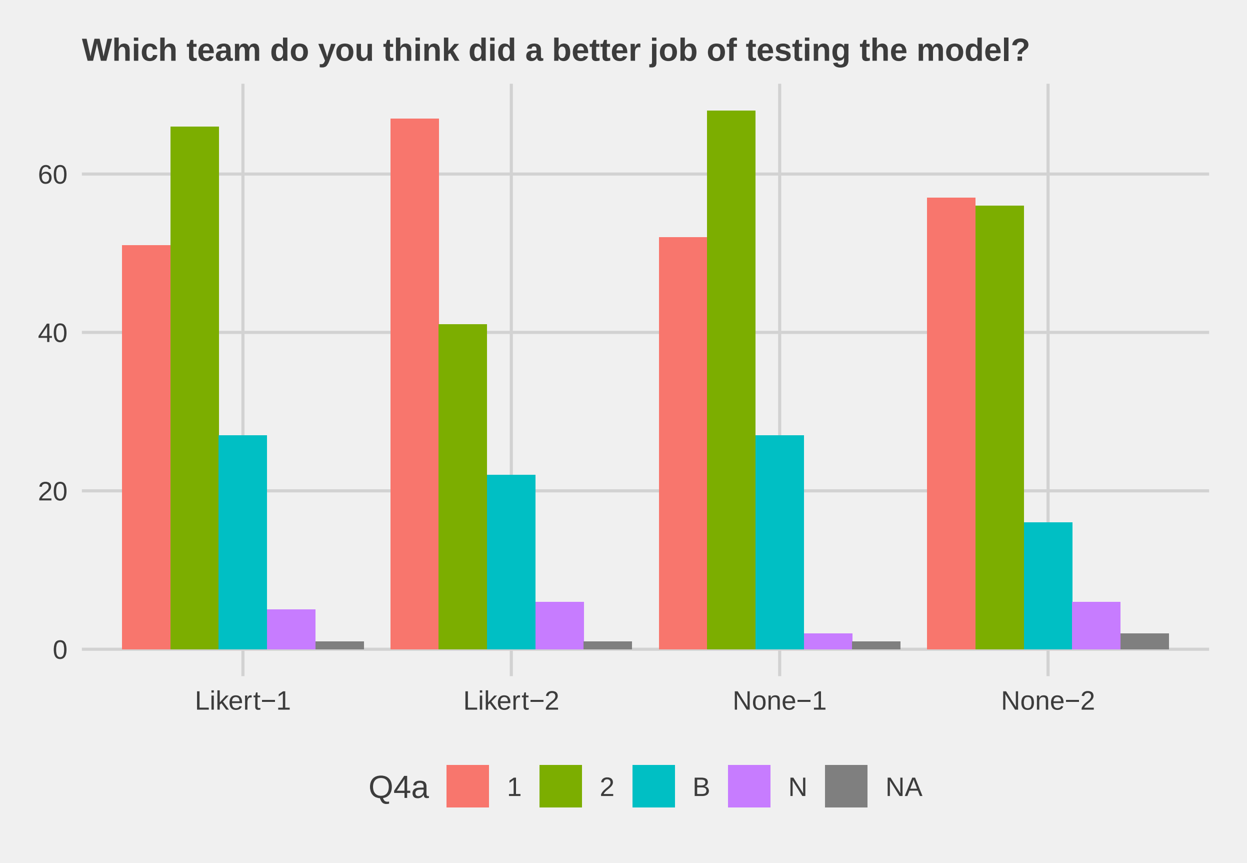
Load and combine students' responses

Analysis

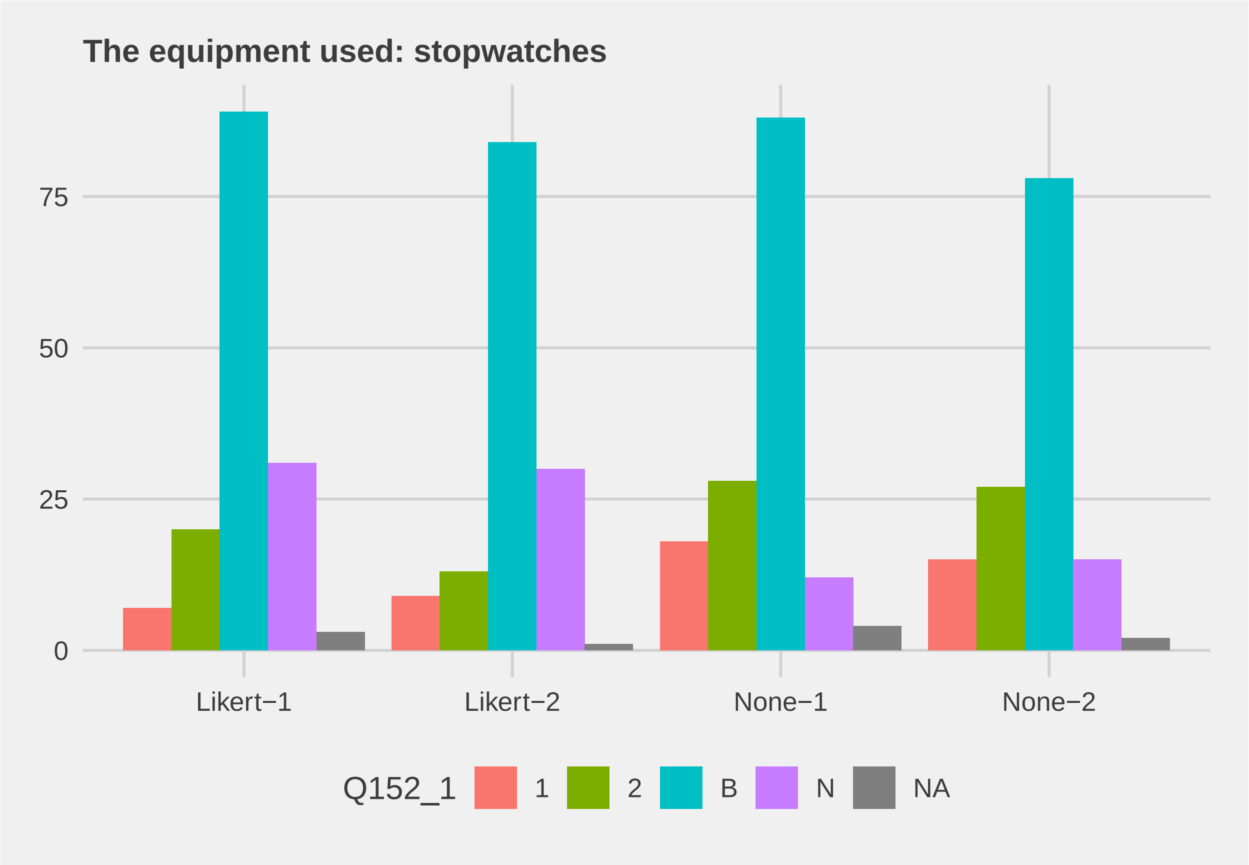
Plots and chi-squared tests

```
##
## Likert-1 Likert-2   None-1   None-2
##      150      137      150      137
```

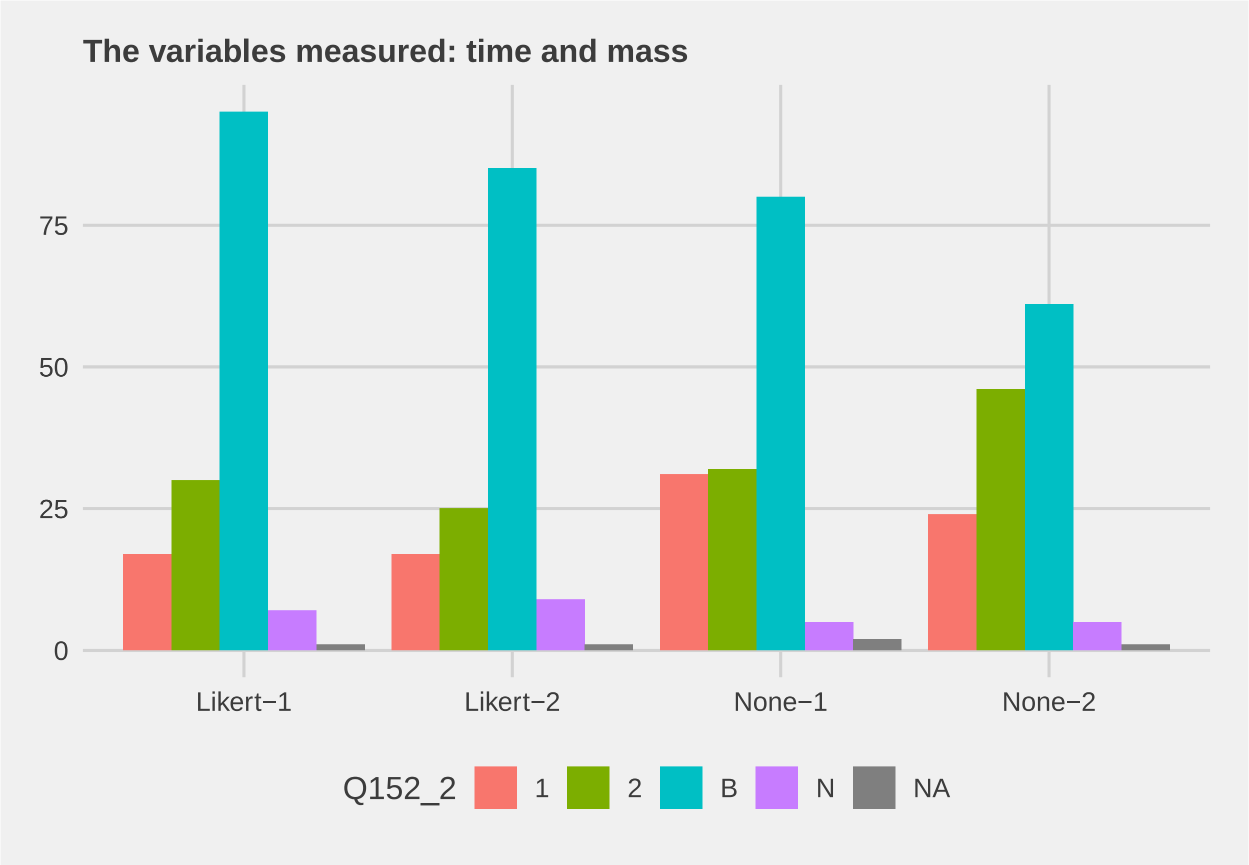
```
## [1] "Which team do you think did a better job of testing the model?"
```



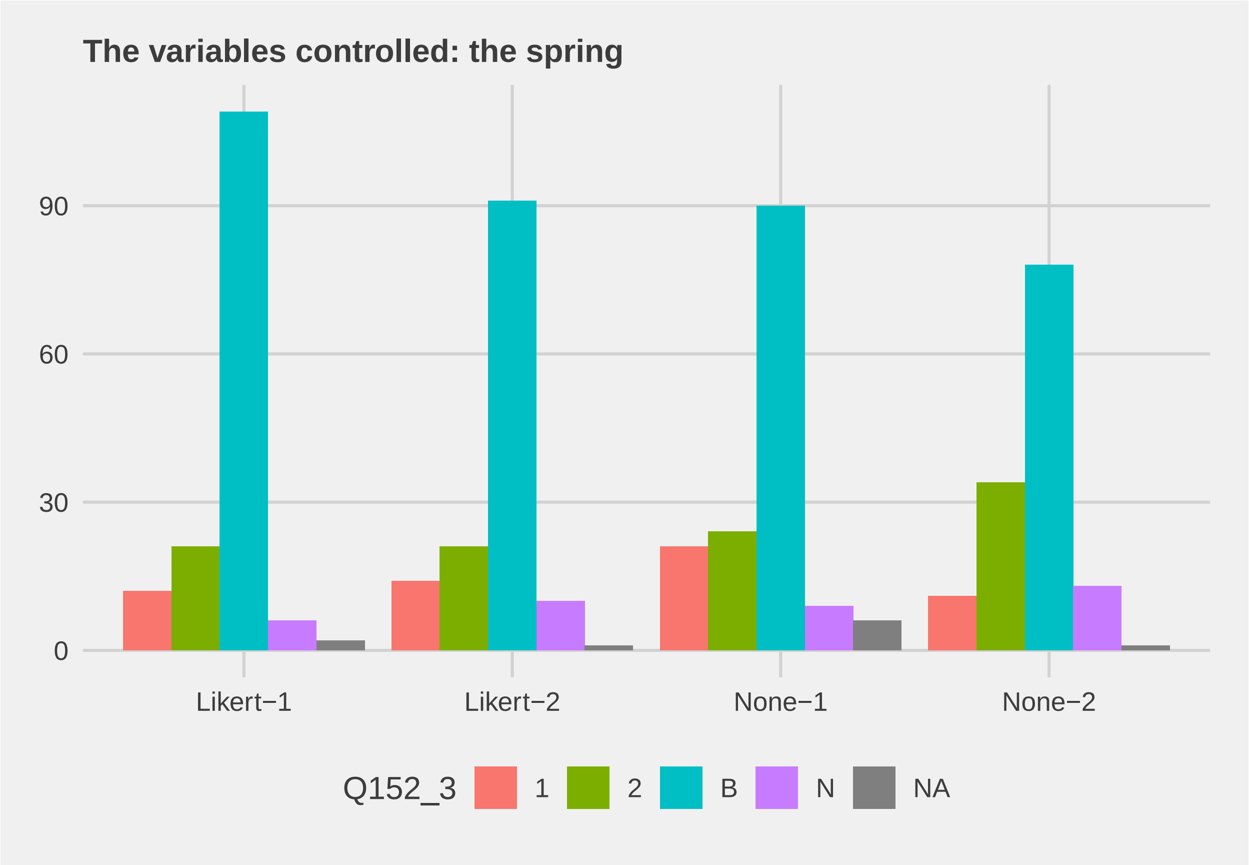
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 15.426, df = 9, p-value = 0.07987
##
## [1] "The equipment used: stopwatches"
```



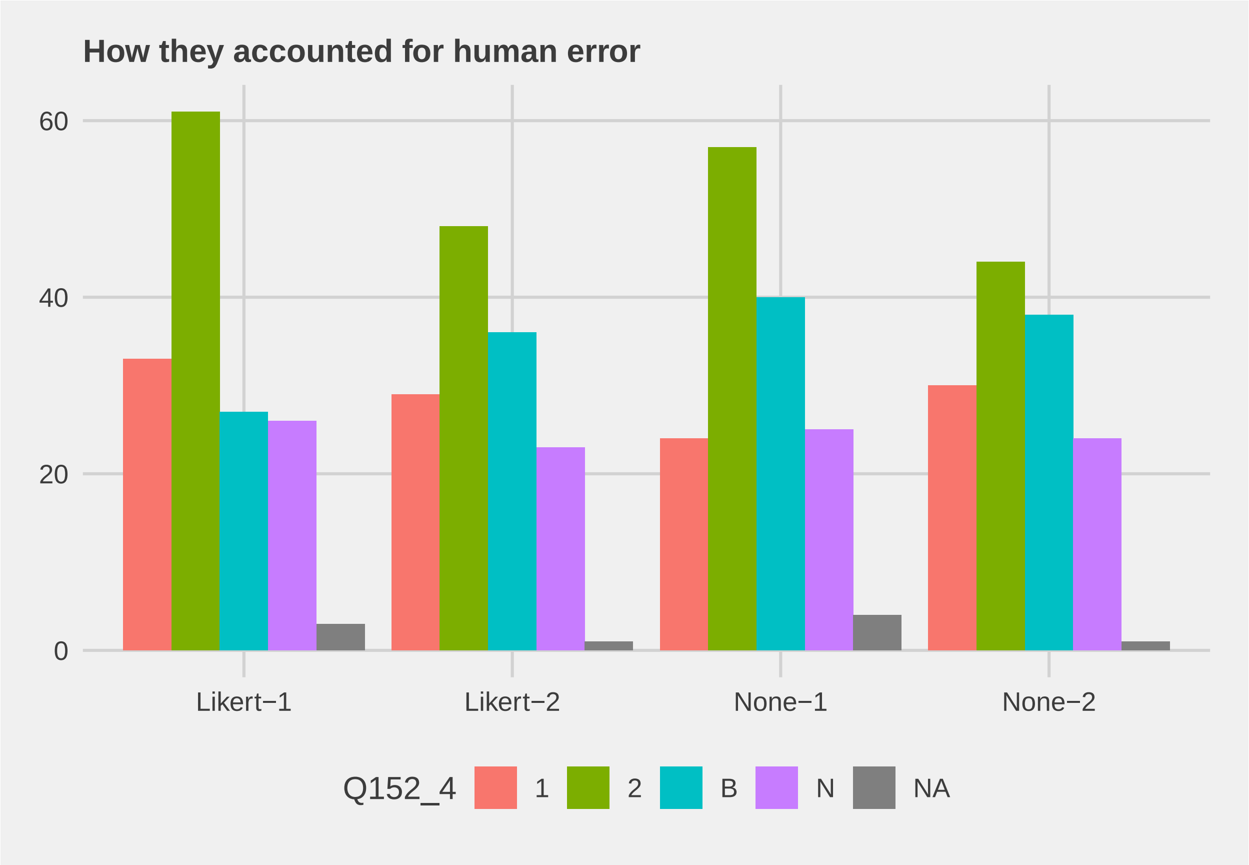
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 26.352, df = 9, p-value = 0.001789
##
## [1] "The variables measured: time and mass"
```


The variables measured: time and mass

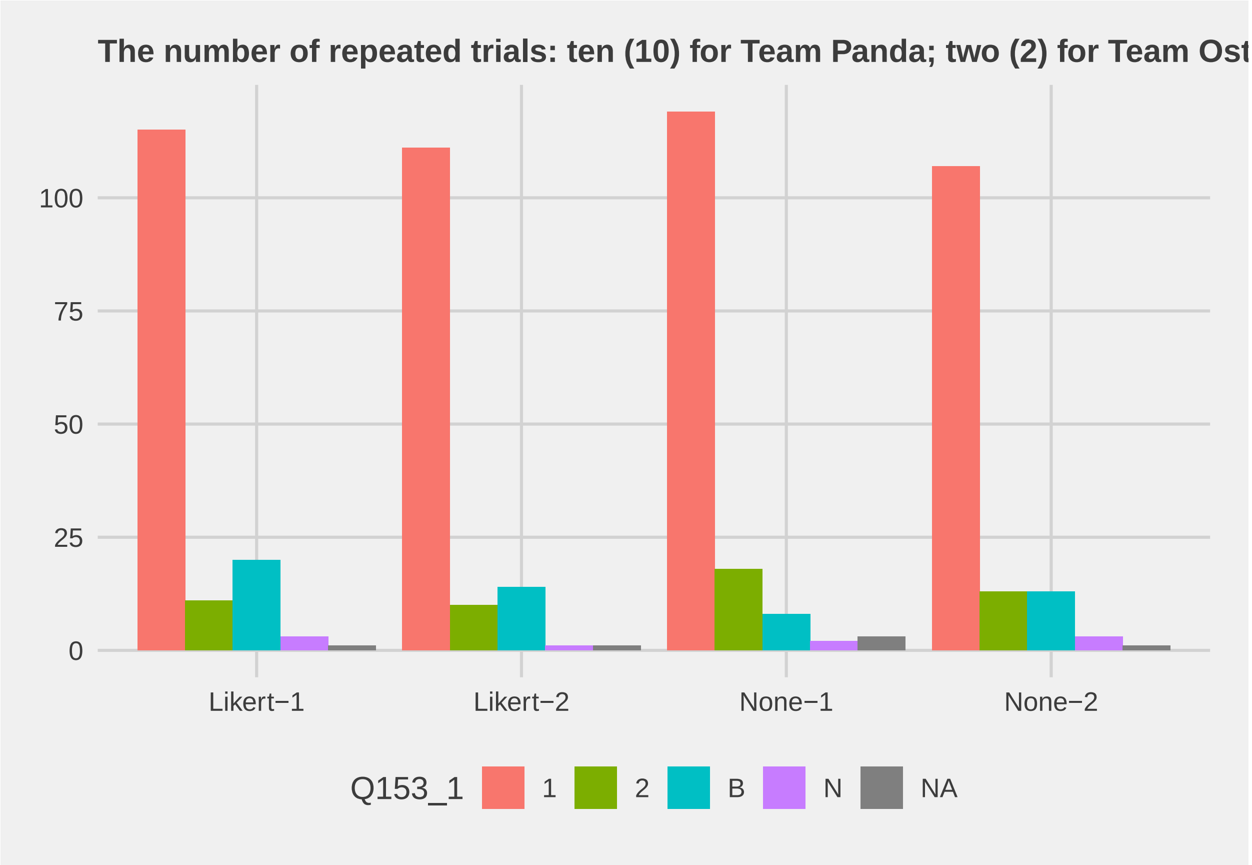
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 21.879, df = 9, p-value = 0.009269
##
## [1] "The variables controlled: the spring"
```



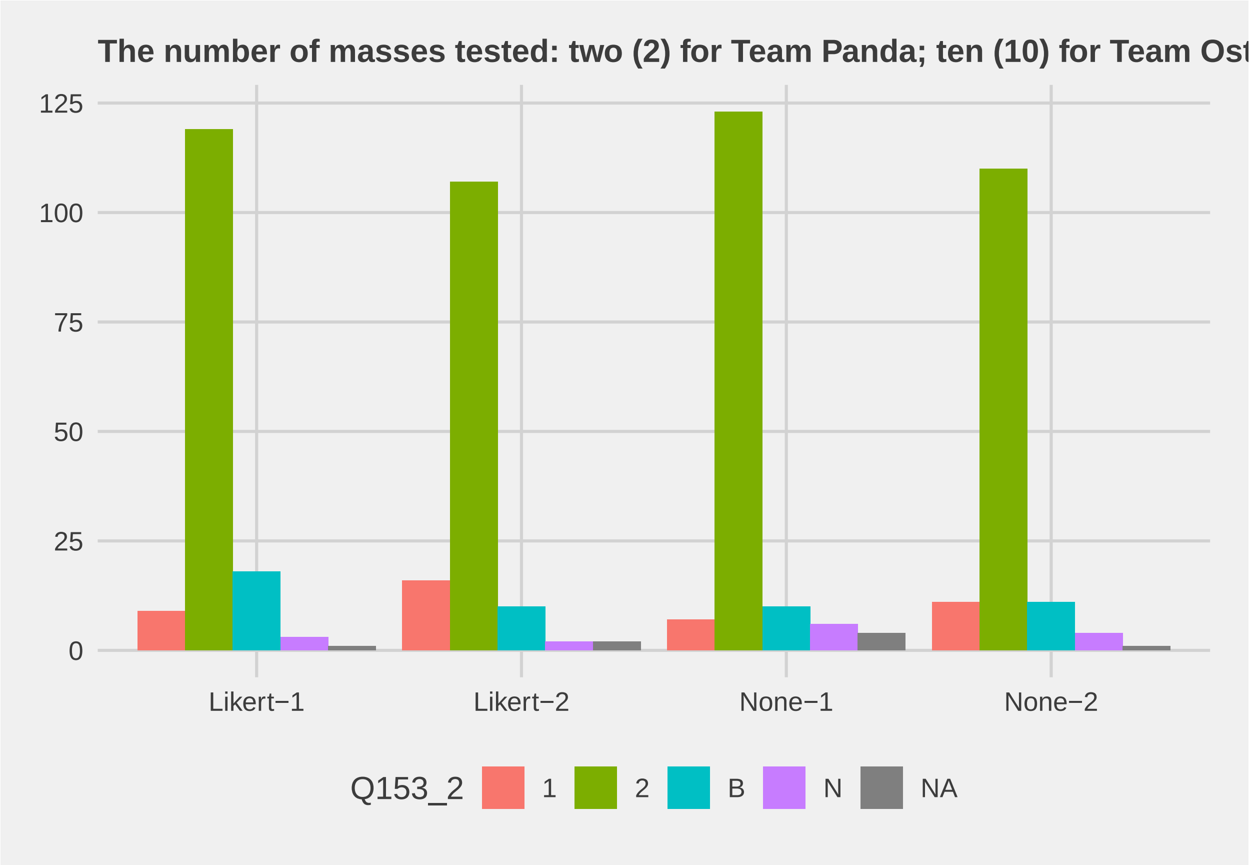
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 15.942, df = 9, p-value = 0.06811
##
## [1] "How they accounted for human error"
```



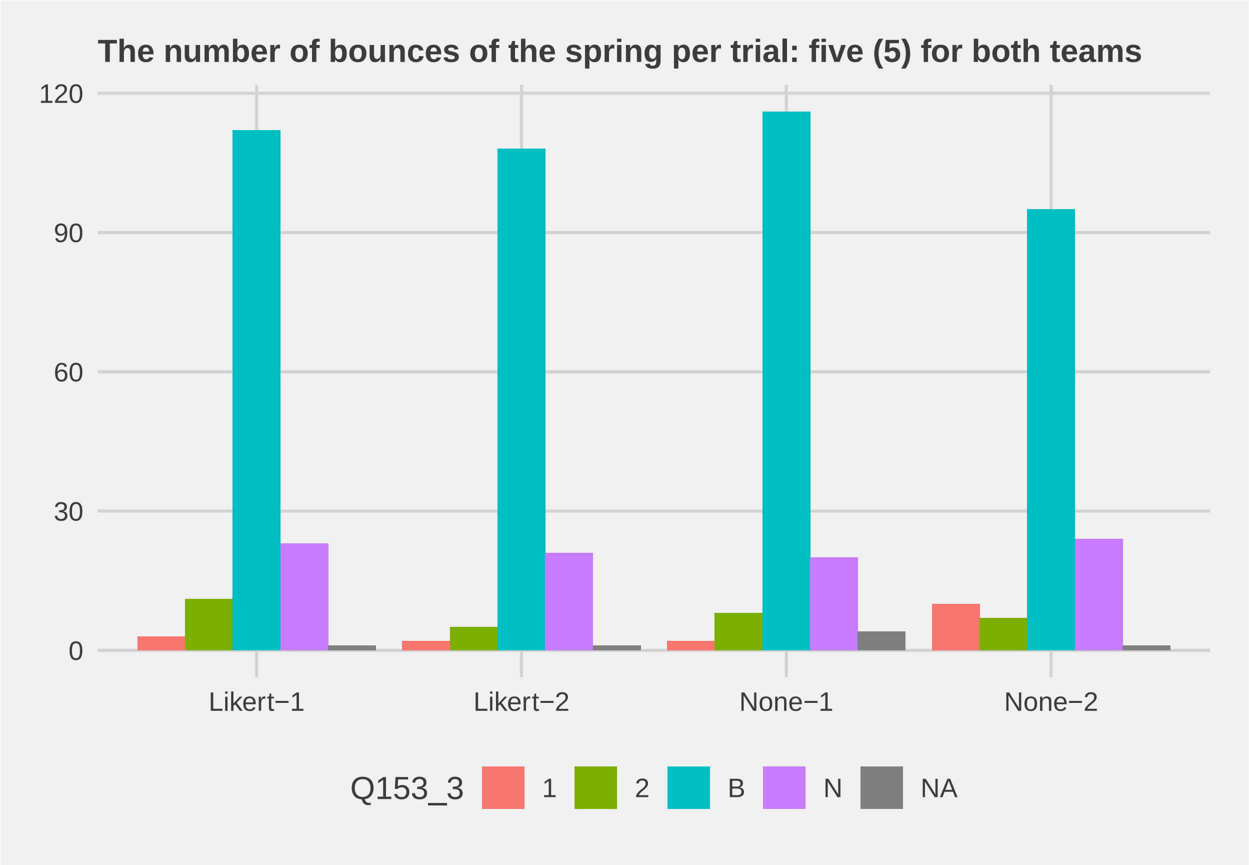
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 7.0587, df = 9, p-value = 0.631
##
## [1] "The number of repeated trials: ten (10) for Team Panda; two (2) for Team Ostrich"
```



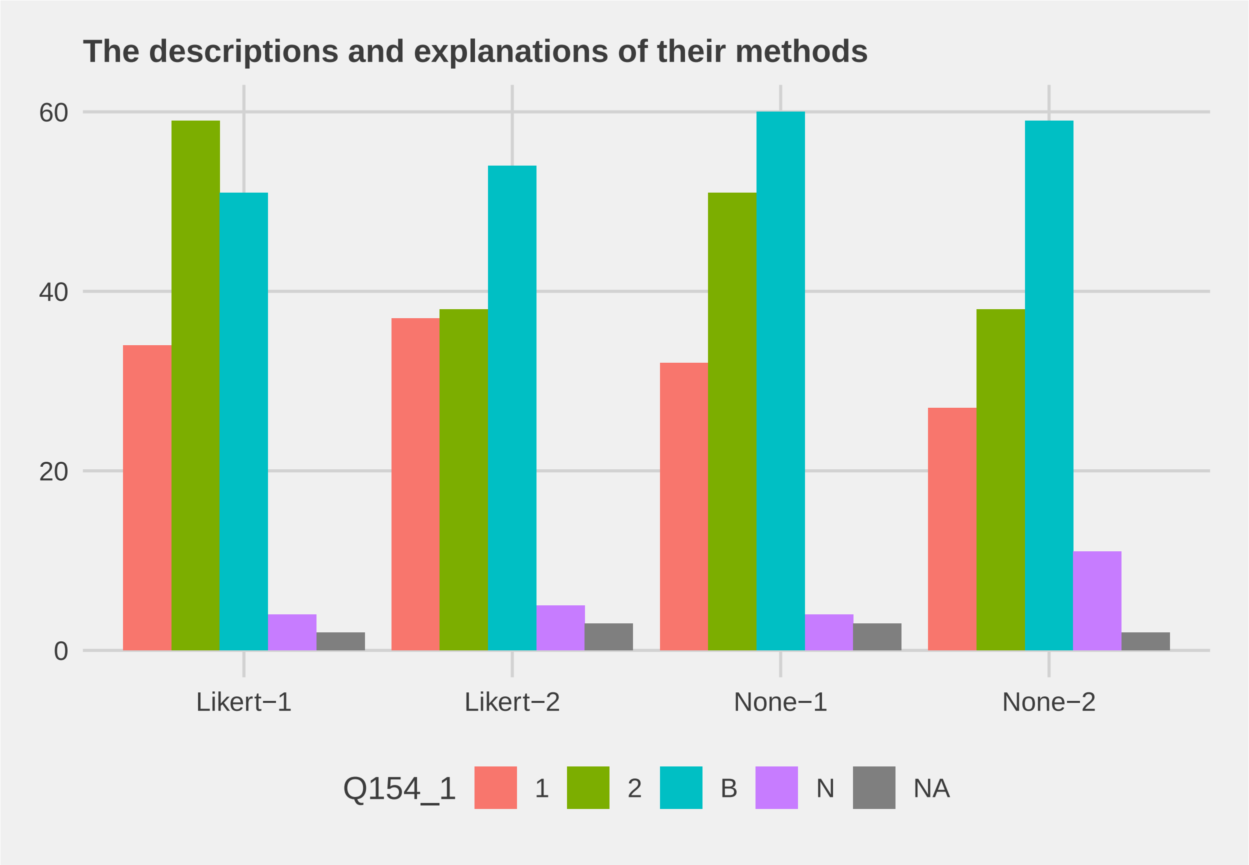
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 8.8854, df = 9, p-value = 0.4479
##
## [1] "The number of masses tested: two (2) for Team Panda; ten (10) for Team Ostrich"
```



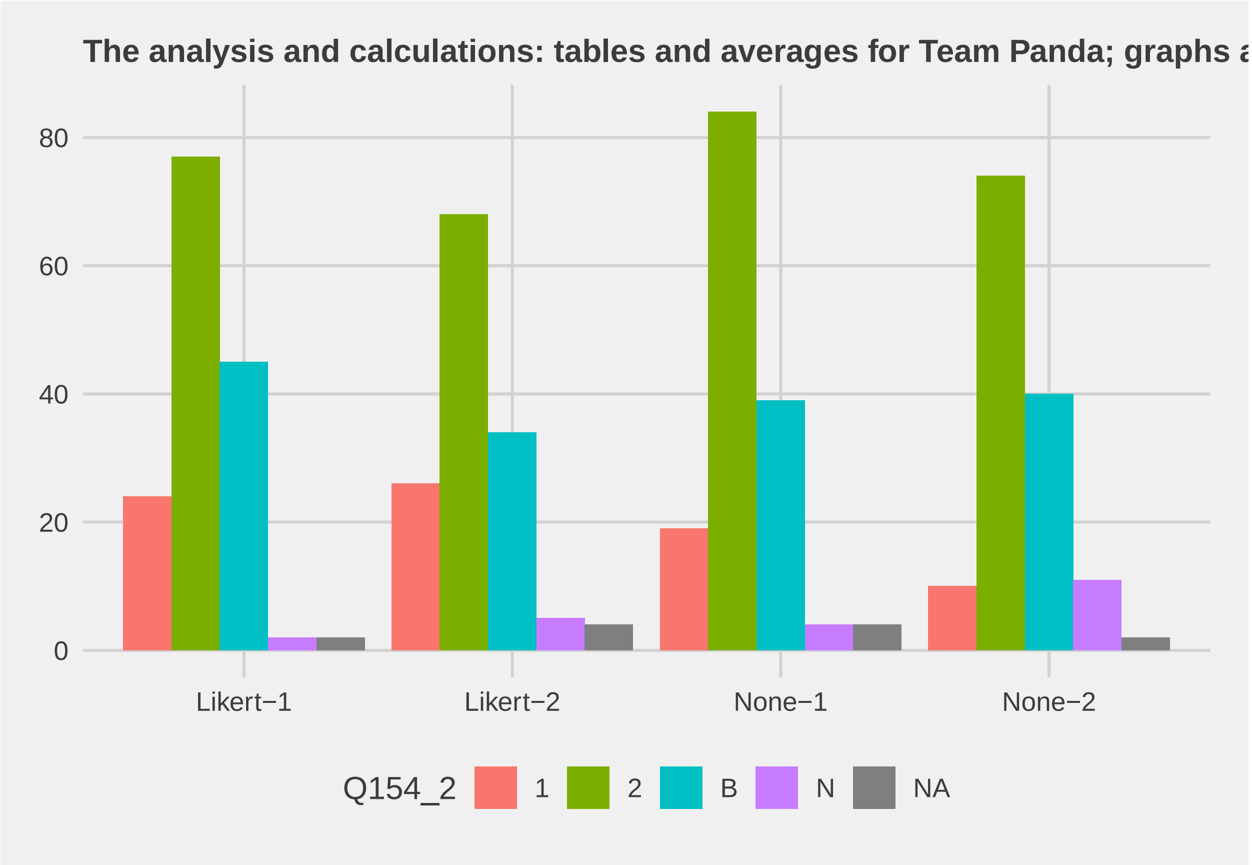
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 10.511, df = 9, p-value = 0.3107
##
## [1] "The number of bounces of the spring per trial: five (5) for both teams"
```



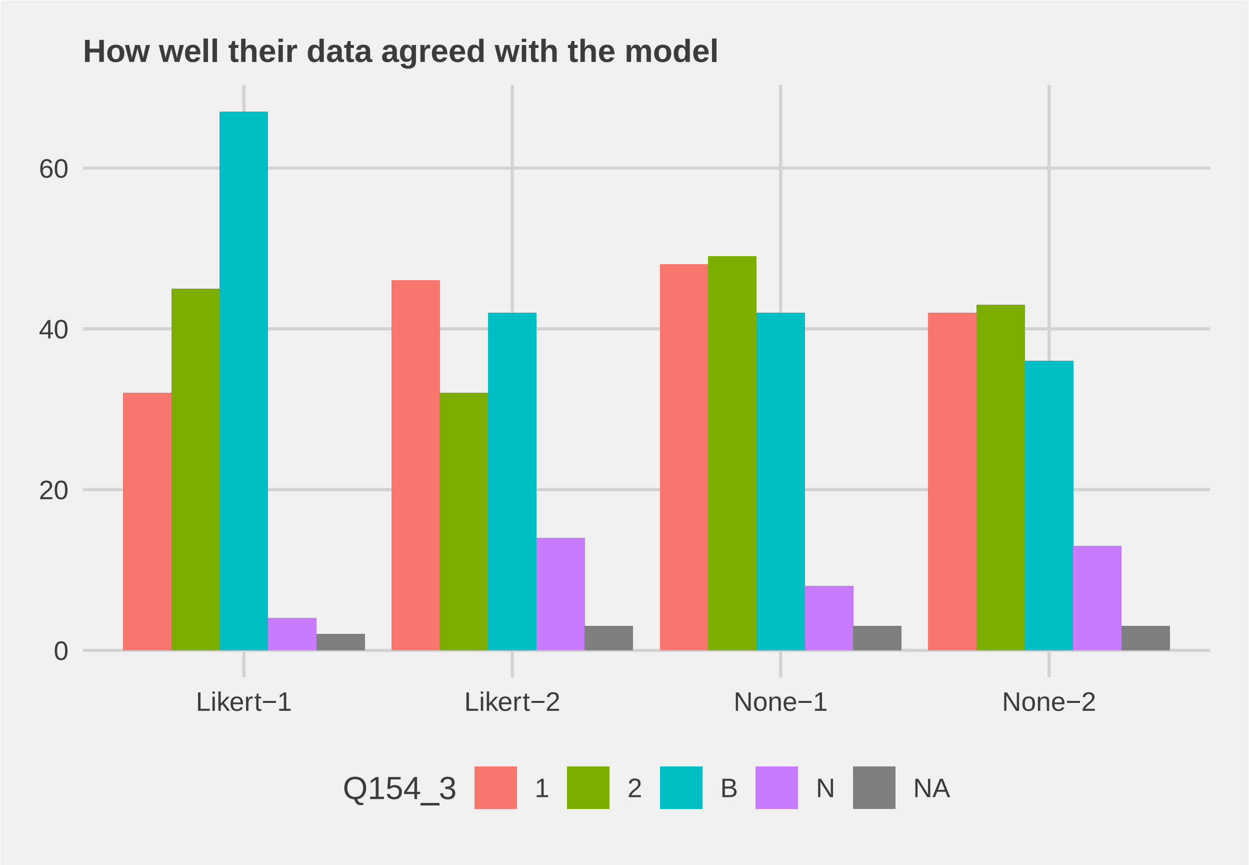
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 15.093, df = 9, p-value = 0.08841
##
## [1] "The descriptions and explanations of their methods"
```



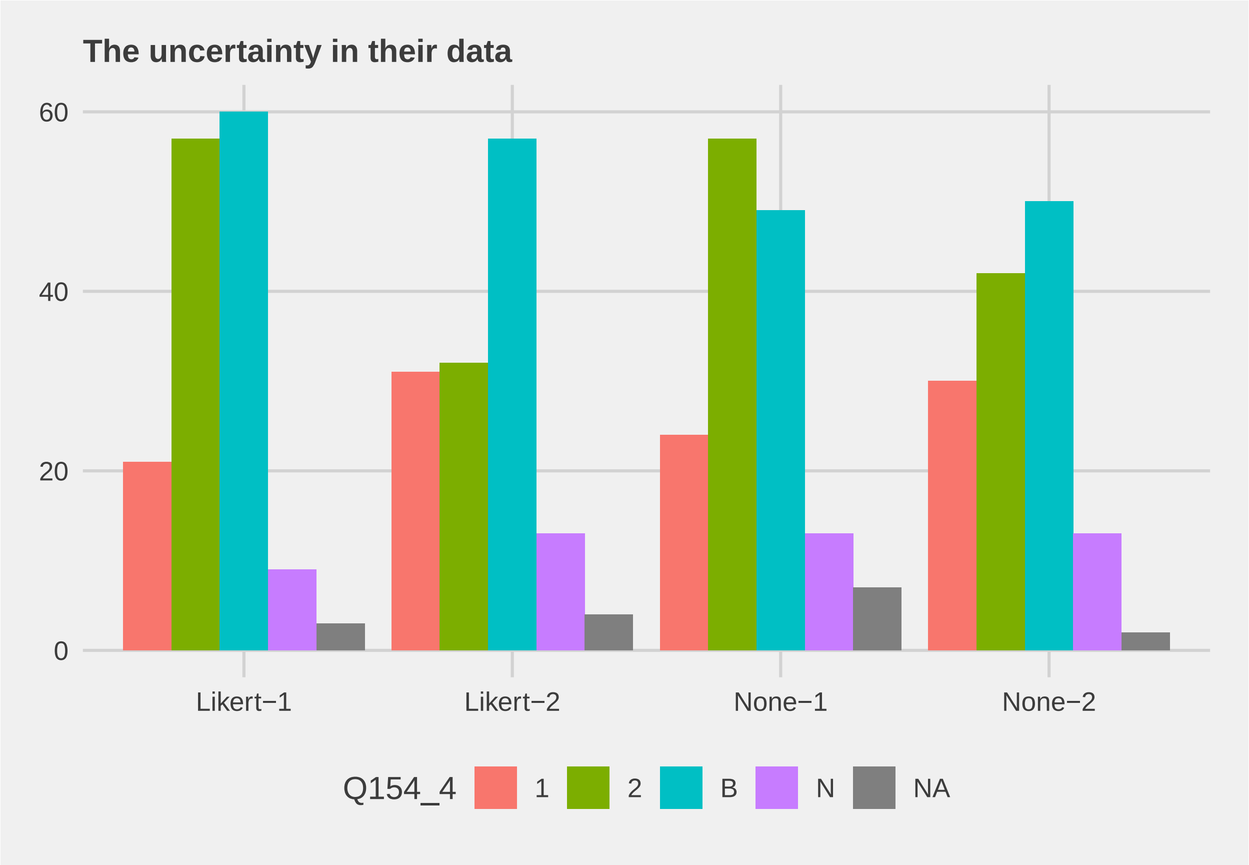
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 14.114, df = 9, p-value = 0.1183
##
## [1] "The analysis and calculations: tables and averages for Team Panda; graphs and fits for Team Ost
```



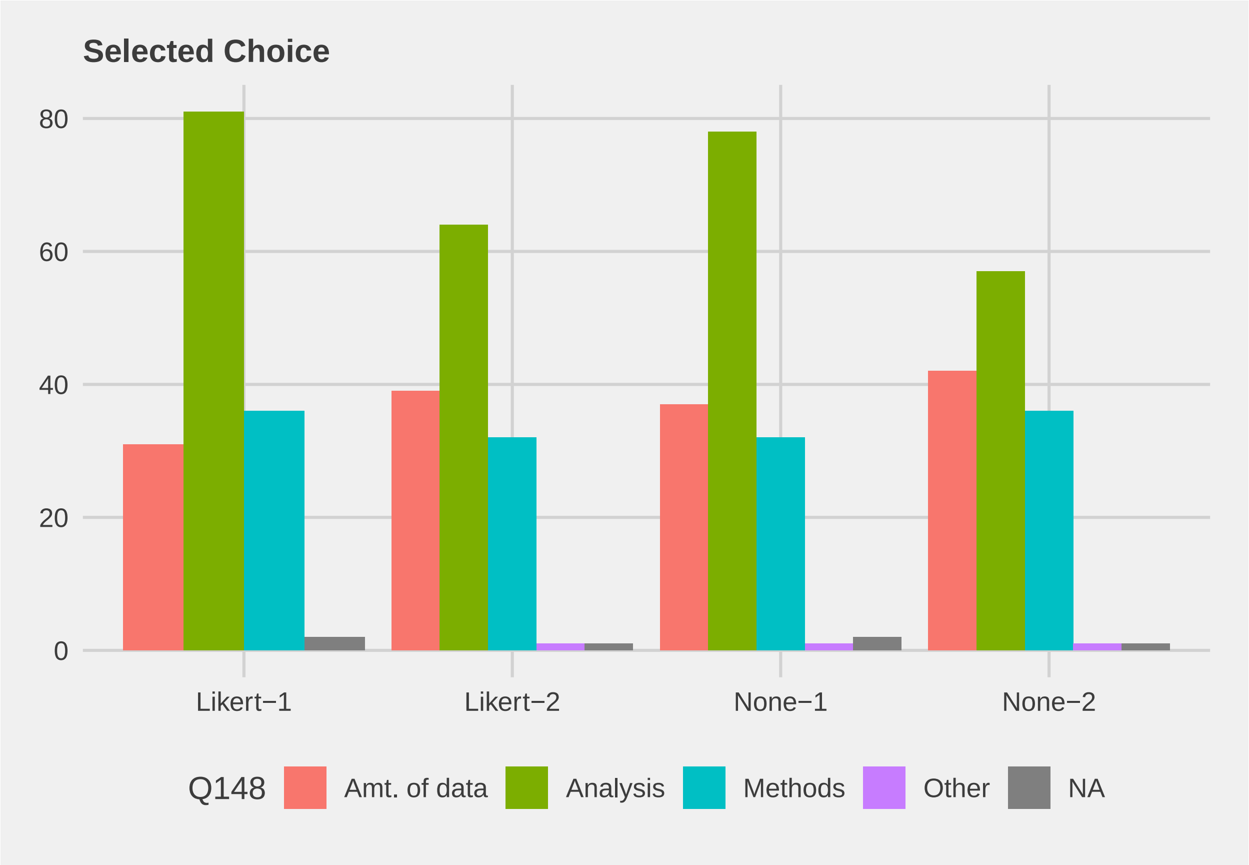
```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 18.353, df = 9, p-value = 0.03129
##
## [1] "How well their data agreed with the model"
```



```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 24.53, df = 9, p-value = 0.003538
##
## [1] "The uncertainty in their data"
```



```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 13.854, df = 9, p-value = 0.1276
##
## [1] "Selected Choice"
```



```
##
##  Pearson's Chi-squared test
##
## data:  df.students$Condition and df.students[, Q]
## X-squared = 7.7486, df = 9, p-value = 0.5597
```

Prefer not to examine all of these items individually because we'll run into multiple comparisons issues and troubles parsing all of this information, but I think it is worth looking at the first and last summary question (Q4a: Which group do you think did a better job? and Q148: What feature was most important for comparing the two teams?) We fail to reject the null hypothesis (at alpha = 0.05) that either distribution of selections differ by condition, but I think there are some trends in both, particularly in the effect of putting Group 2 first.

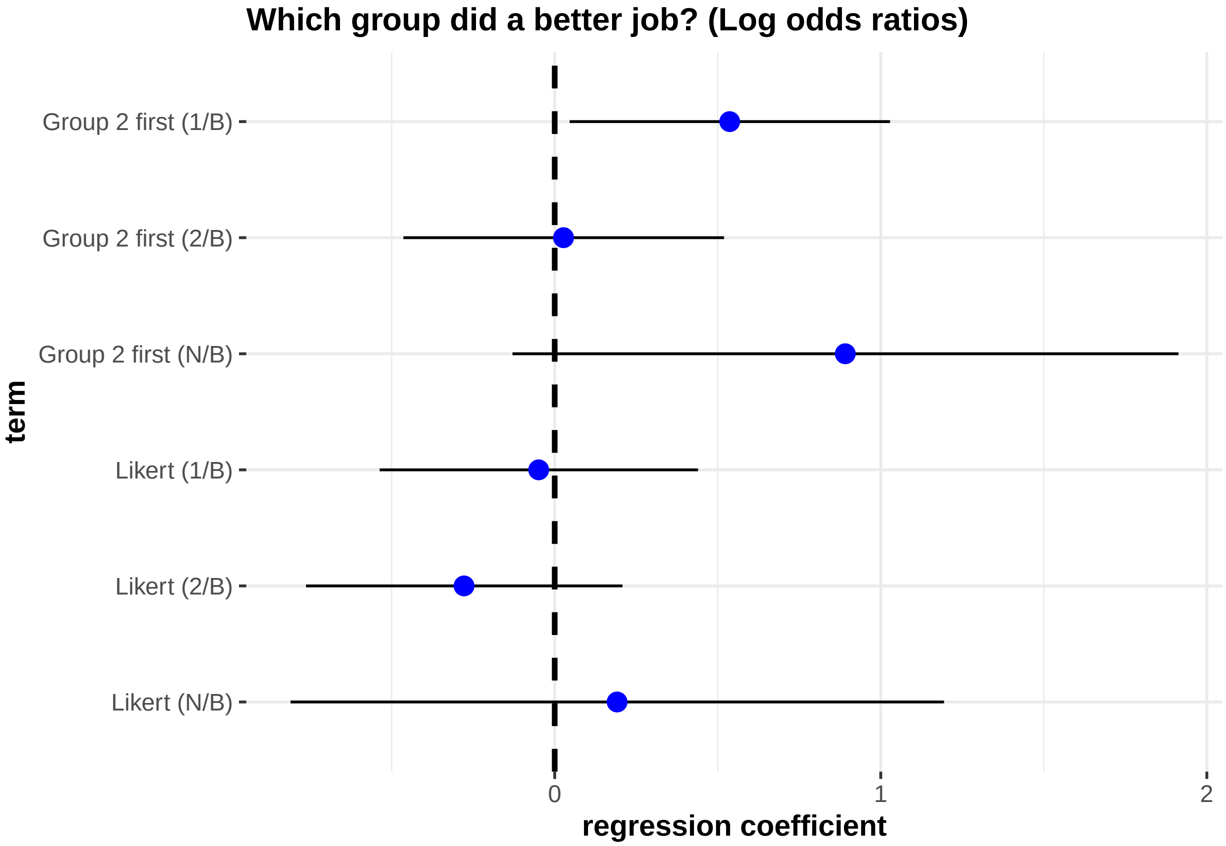
For Q4a, more students look to pick Group 1 and less pick Group 2 when shown Group 2 first. Though less apparent, fewer students identify "Analysis" as being important when shown Group 2 first.

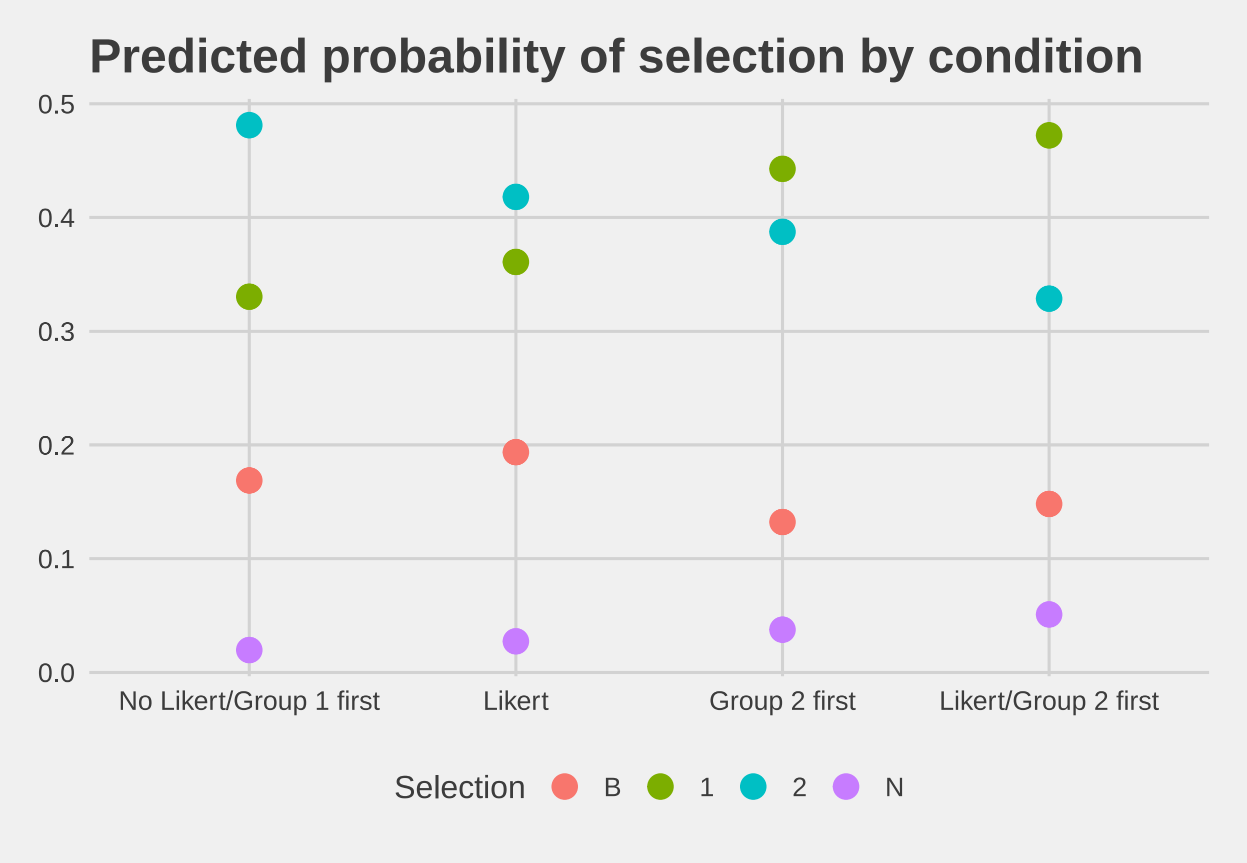
Multinomial model of "Who did better?"

```
## # weights:  16 (9 variable)
## initial  value 788.801491
## iter  10 value 642.900033
## final  value 642.443772
## converged
```

```
## Note: The argument `statistic` must be specified.
##  Skipping labels with statistical details.
```

##


Which group did a better job? (Log odds ratios)
Group 2 first (1/B)
Group 2 first (2/B)
Group 2 first (N/B)
Likert (1/B)
Likert (2/B)
Likert (N/B)
term
0
1
2
regression coefficient



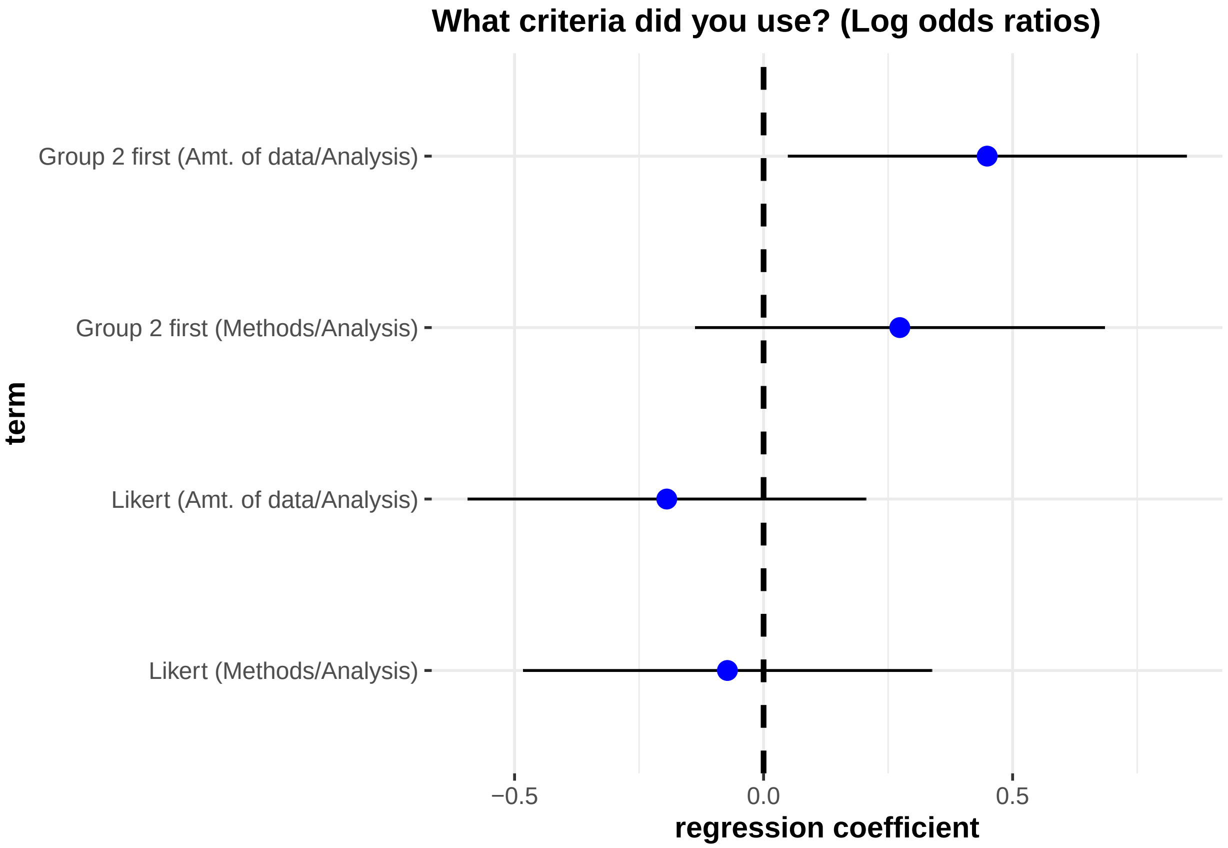
This multinomial illustrates these effects for Q4a. Showing the Likert items in the survey has little to no effect on students' responses to Q4a, but a greater proportion of students select Group 1 and N when shown Group 2 first. I think its easier to interpret the size of this effect by looking at the expected proportions because we had a 2x2 design. The fraction of students selecting Group 2 decreases from almost 0.5 to just below 0.4 when shown Group 2 first. The fraction selecting Group 1 conversely increases by almost 10 percentage points. These effects are considerably smaller for the Likert condition.

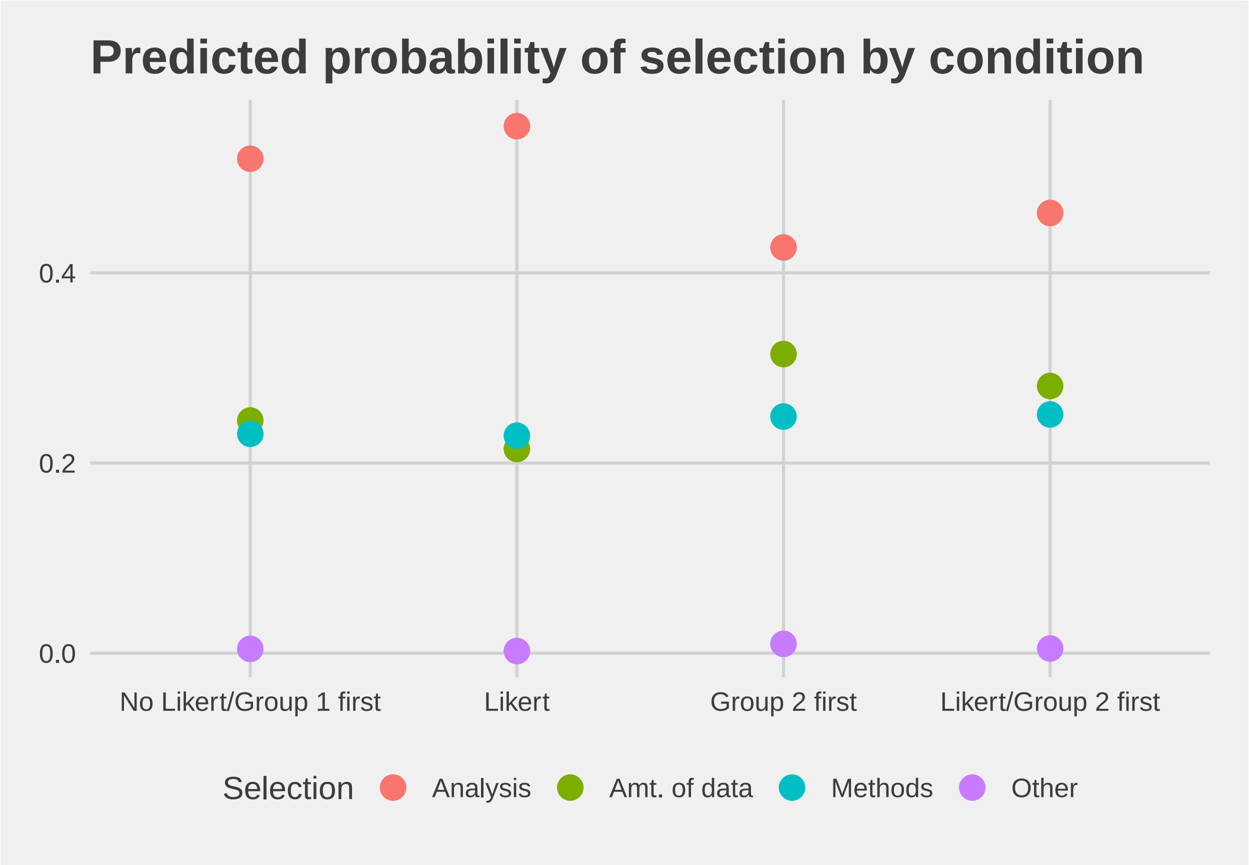
Multinomial model of "What was most important?"

```
## # weights:  16 (9 variable)
## initial  value 787.415197
## iter  10 value 613.928600
## iter  20 value 604.527289
## iter  30 value 604.135967
## final  value 604.135932
## converged
```

```
## Note: The argument `statistic` must be specified.
##  Skipping labels with statistical details.
```

```
##
```

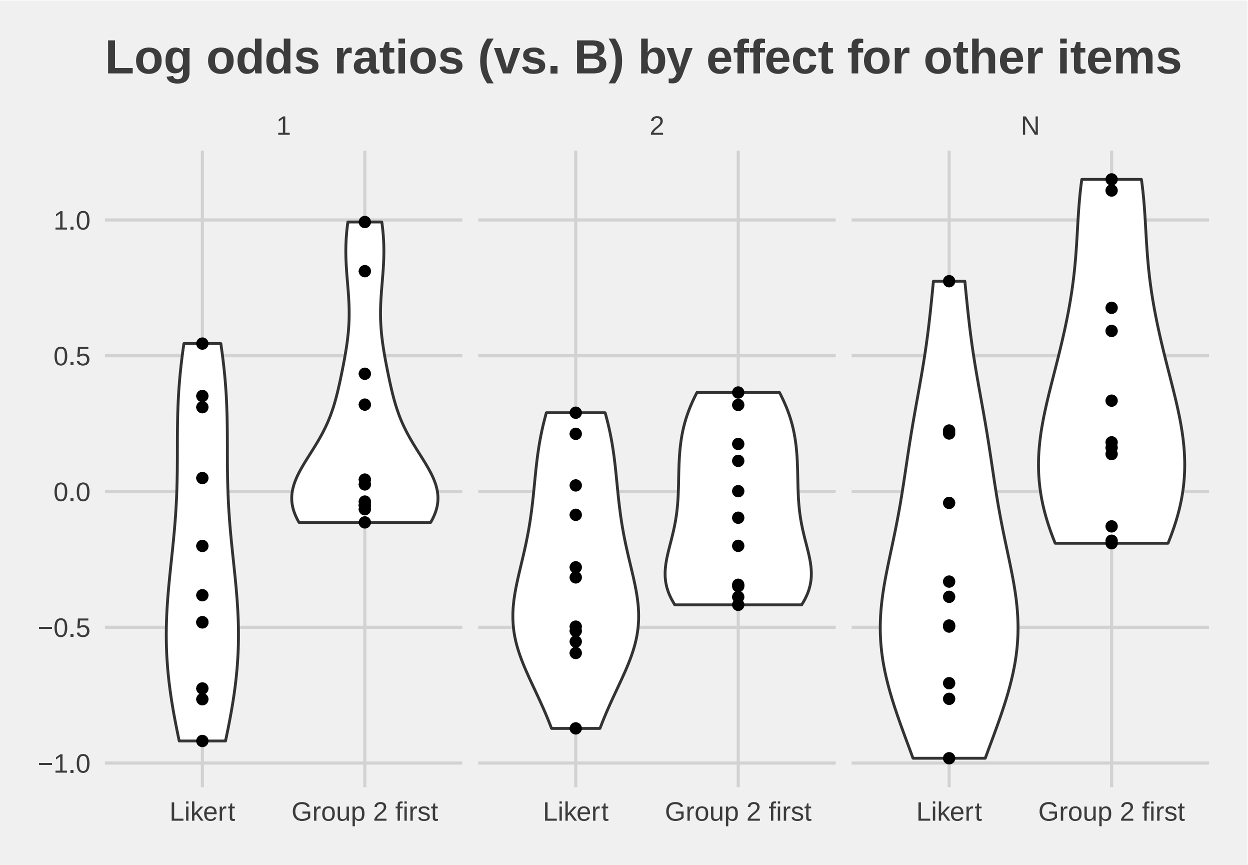

What criteria did you use? (Log odds ratios)
term
Group 2 first (Amt. of data/Analysis)
Group 2 first (Methods/Analysis)
Likert (Amt. of data/Analysis)
Likert (Methods/Analysis)
−0.5
0.0
0.5
regression coefficient



The effects for Q148 are smaller overall, but we again see that swapping the groups has a larger effect than including the Likert items, which is about zero.

Other multinomial models

```
## # weights:  16 (9 variable)
## initial  value 781.870020
## iter  10 value 607.203810
## final  value 606.337384
## converged
## # weights:  16 (9 variable)
## initial  value 788.801491
## iter  10 value 614.498132
## final  value 613.424866
## converged
## # weights:  16 (9 variable)
## initial  value 781.870020
## iter  10 value 558.641163
## final  value 558.334694
## converged
## # weights:  16 (9 variable)
## initial  value 783.256314
## iter  10 value 758.670987
## final  value 756.173007
## converged
## # weights:  16 (9 variable)
## initial  value 787.415197
## iter  10 value 403.567428
## final  value 390.534227
## converged
## # weights:  16 (9 variable)
## initial  value 784.642608
## iter  10 value 387.856477
## iter  20 value 376.965955
## iter  20 value 376.965955
## final  value 376.965955
## converged
## # weights:  16 (9 variable)
## initial  value 786.028903
## iter  10 value 468.470591
## iter  20 value 427.723437
## final  value 427.723422
## converged
## # weights:  16 (9 variable)
## initial  value 781.870020
## iter  10 value 674.241579
## final  value 673.889334
## converged
## # weights:  16 (9 variable)
## initial  value 779.097431
## iter  10 value 612.067304
## final  value 606.405199
## converged
## # weights:  16 (9 variable)
## initial  value 780.483725
## iter  10 value 707.514013
## final  value 706.799655
## converged
## # weights:  16 (9 variable)
## initial  value 773.552254
## iter  10 value 697.129005
## final  value 696.879441
## converged
```

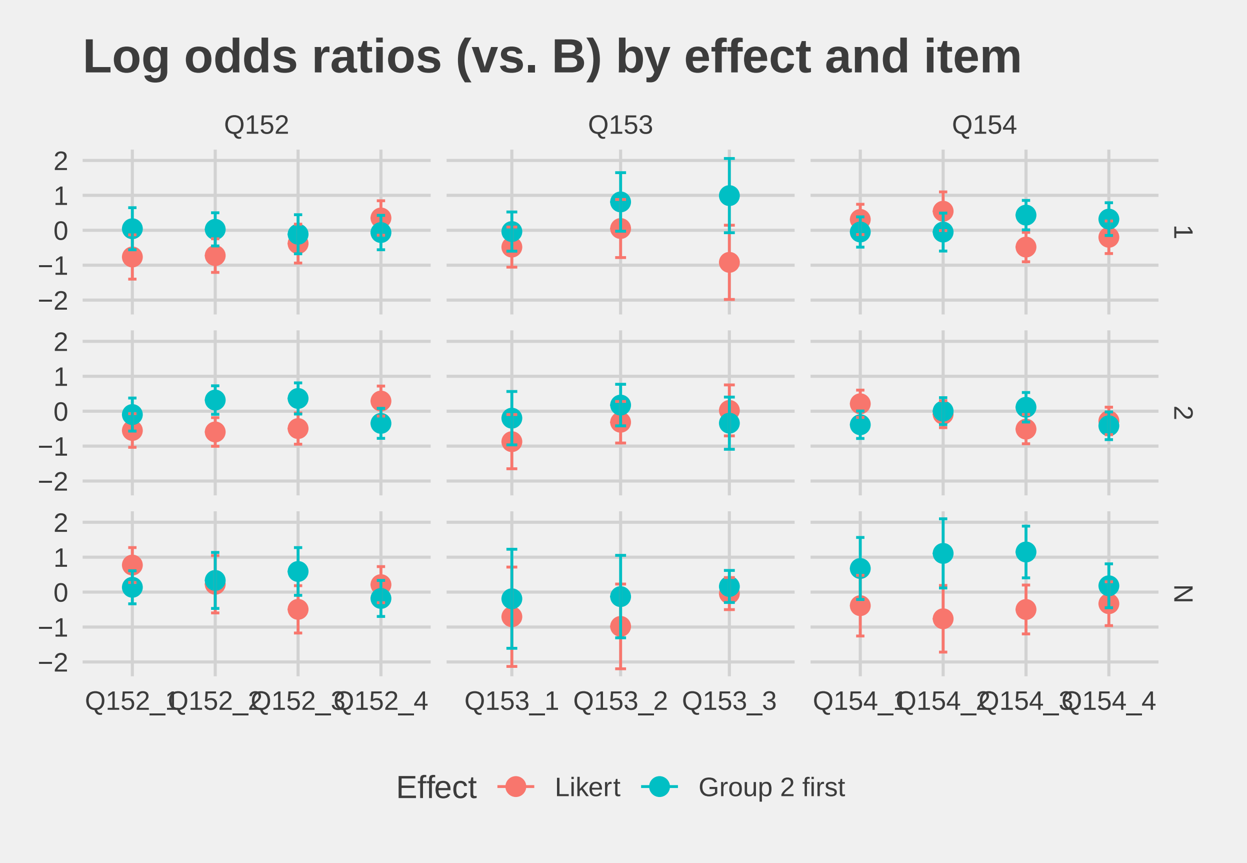


I've shown all the other 11 items here without distinguishing between items. A model was fit separately for each item. I think a couple things stand out. First, the Likert items have more variable effects on the summary items, but are generally negative (relative to B), indicating that including the Likert items increases the fraction of students that select B (as was the case with Q4a). Its worth keeping in mind that these effects are small and the error bars are quite large (but not shown for clarity, see below).

Putting Group 2 first, conversely, generally increases the fraction of students that say Group 1 or neither group did well. The ratio of students selecting (Group 2/Both) remains more or less constant across all items, however.

Disentangling collection of other multiple choice items

```
##                                                             t.info...Questions.2..length.Questions...
## Q152_1                                                                     The equipment used: stopwa
## Q152_2                                                               The variables measured: time an
## Q152_3                                                                The variables controlled: the s
## Q152_4                                                                  How they accounted for human
## Q153_1                    The number of repeated trials: ten (10) for Team Panda; two (2) for Team Os
## Q153_2                      The number of masses tested: two (2) for Team Panda; ten (10) for Team Os
## Q153_3                              The number of bounces of the spring per trial: five (5) for both
## Q154_1                                                  The descriptions and explanations of their mo
## Q154_2 The analysis and calculations: tables and averages for Team Panda; graphs and fits for Team Os
## Q154_3                                                           How well their data agreed with the
## Q154_4                                                                       The uncertainty in thei
```



This plot extends on the above plot and separates effects by item and includes error bars.

Aggregate multinomial model

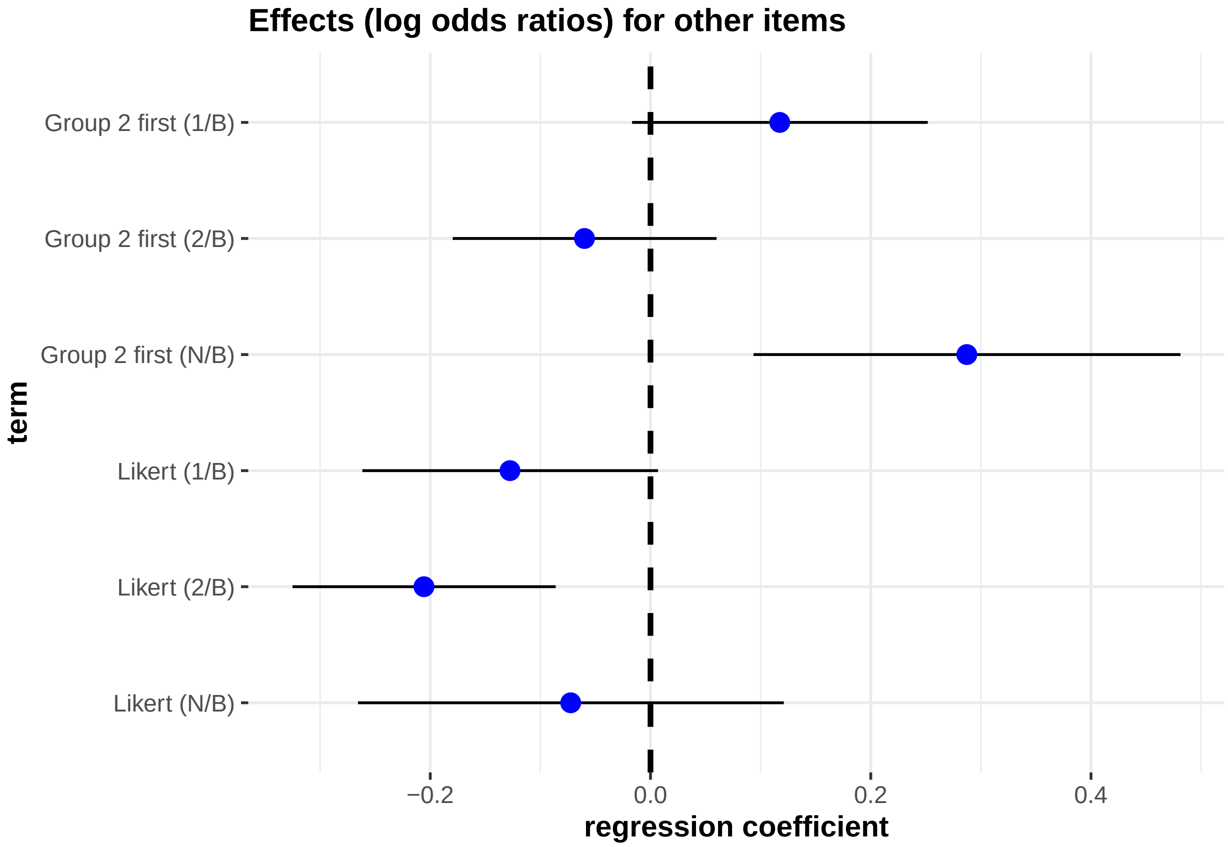
```
## # weights:  16 (9 variable)
## initial  value 8608.887983
## iter  10 value 7876.904744
## final  value 7804.658878
## converged

## # weights:  40 (27 variable)
## initial  value 8608.887983
## iter  10 value 7880.158904
## iter  20 value 7672.189938
## iter  30 value 7615.381174
## final  value 7611.869089
## converged

## # weights:  136 (99 variable)
## initial  value 8608.887983
## iter  10 value 7453.658941
## iter  20 value 6941.635339
## iter  30 value 6597.017607
## iter  40 value 6491.798650
## iter  50 value 6425.800491
## iter  60 value 6415.099513
## iter  70 value 6413.632284
## iter  80 value 6413.468775
## final  value 6413.467213
## converged
```

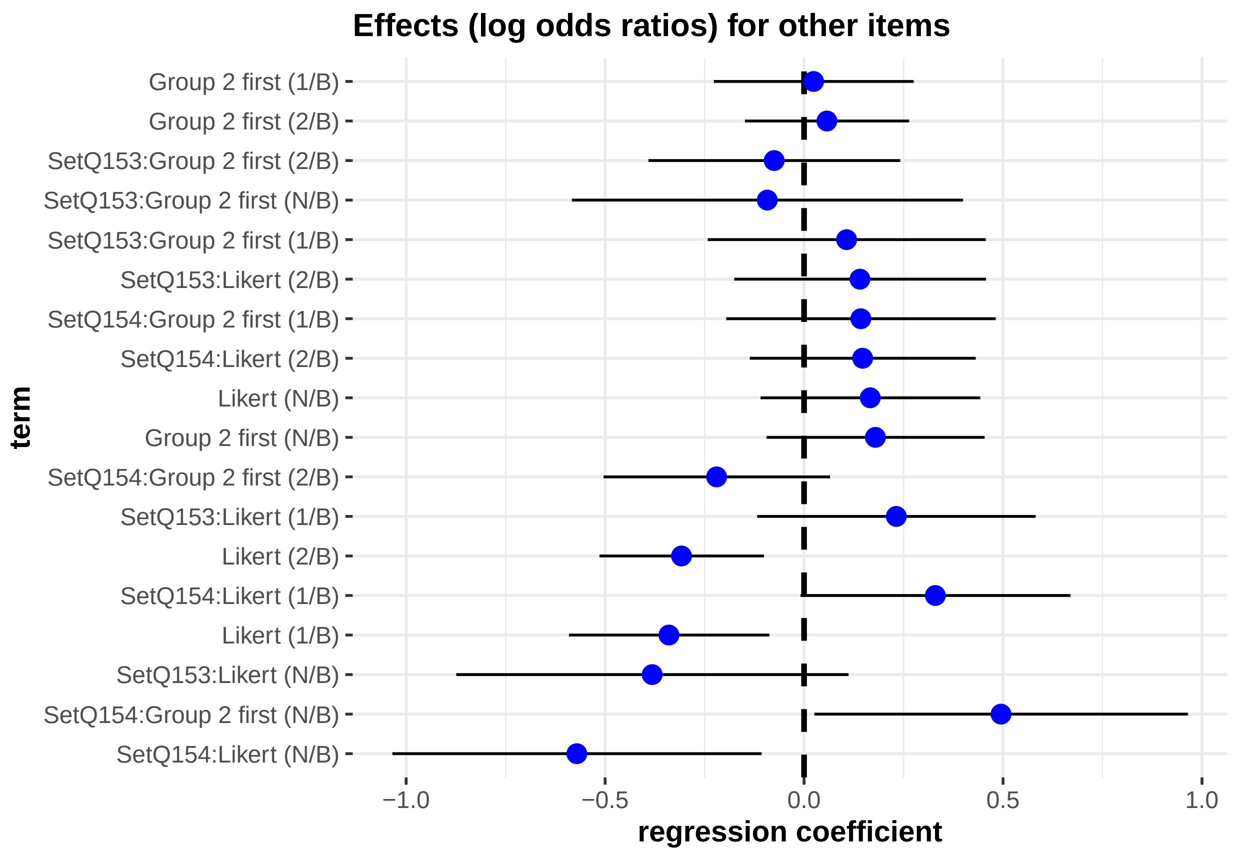
```
## Note: The argument `statistic` must be specified.
##  Skipping labels with statistical details.
```

```
##
```



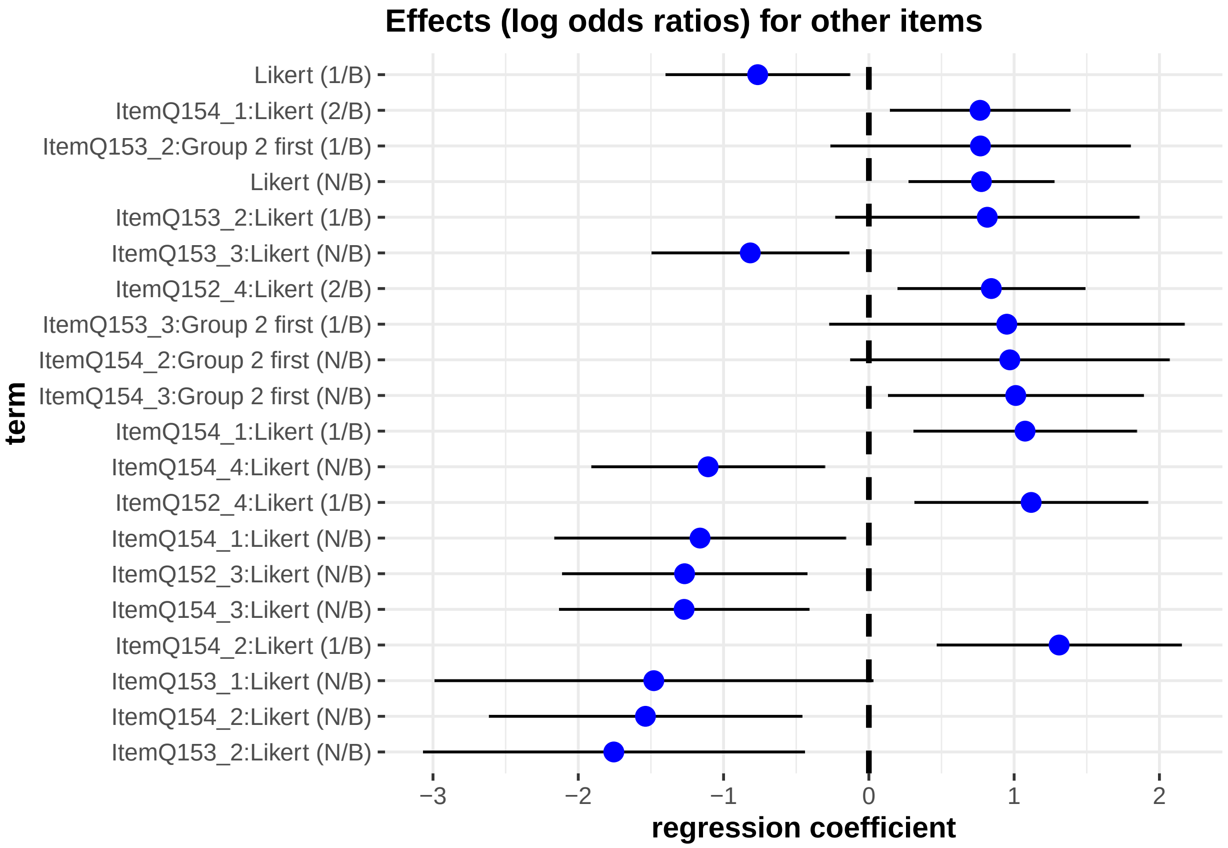
```
## Note: The argument `statistic` must be specified.
##  Skipping labels with statistical details.
```

```
##
```



```
## Note: The argument `statistic` must be specified.
##  Skipping labels with statistical details.
```

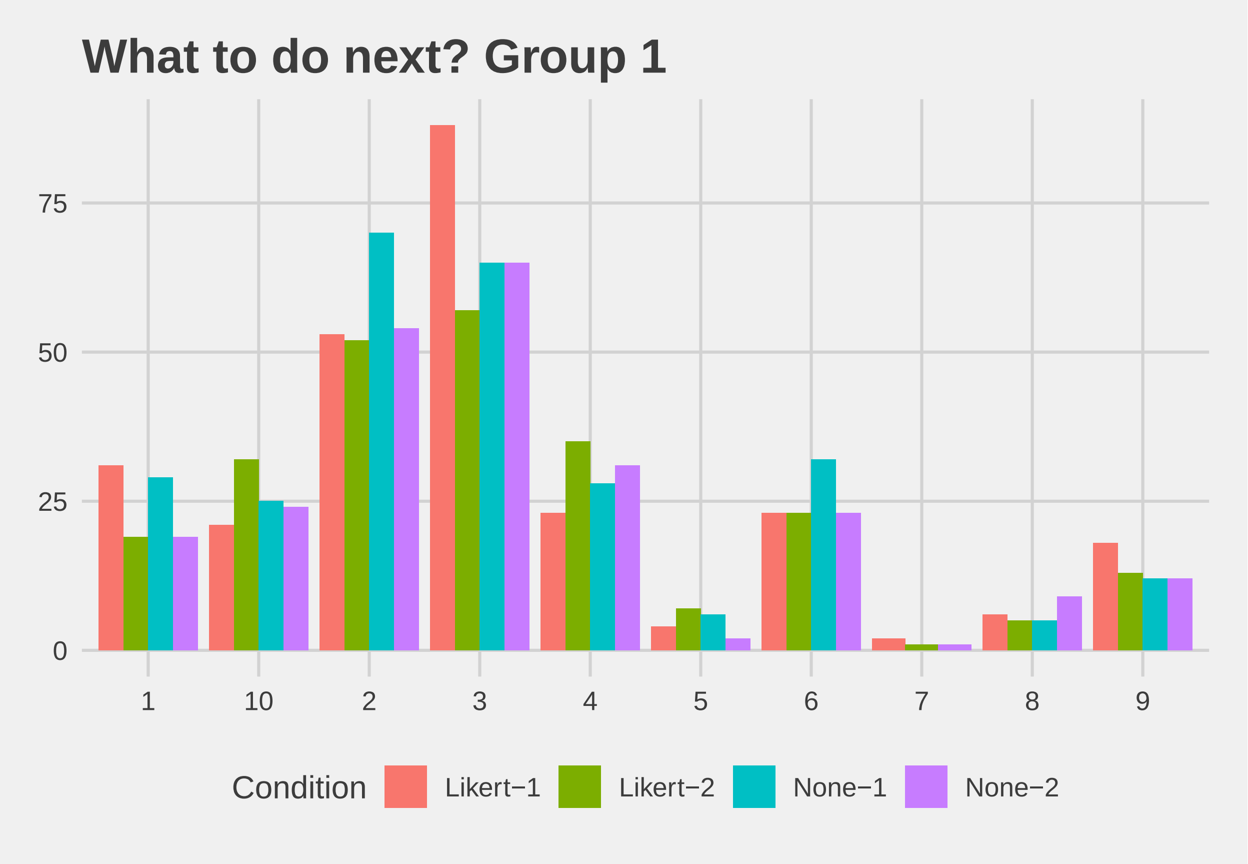
```
##
```



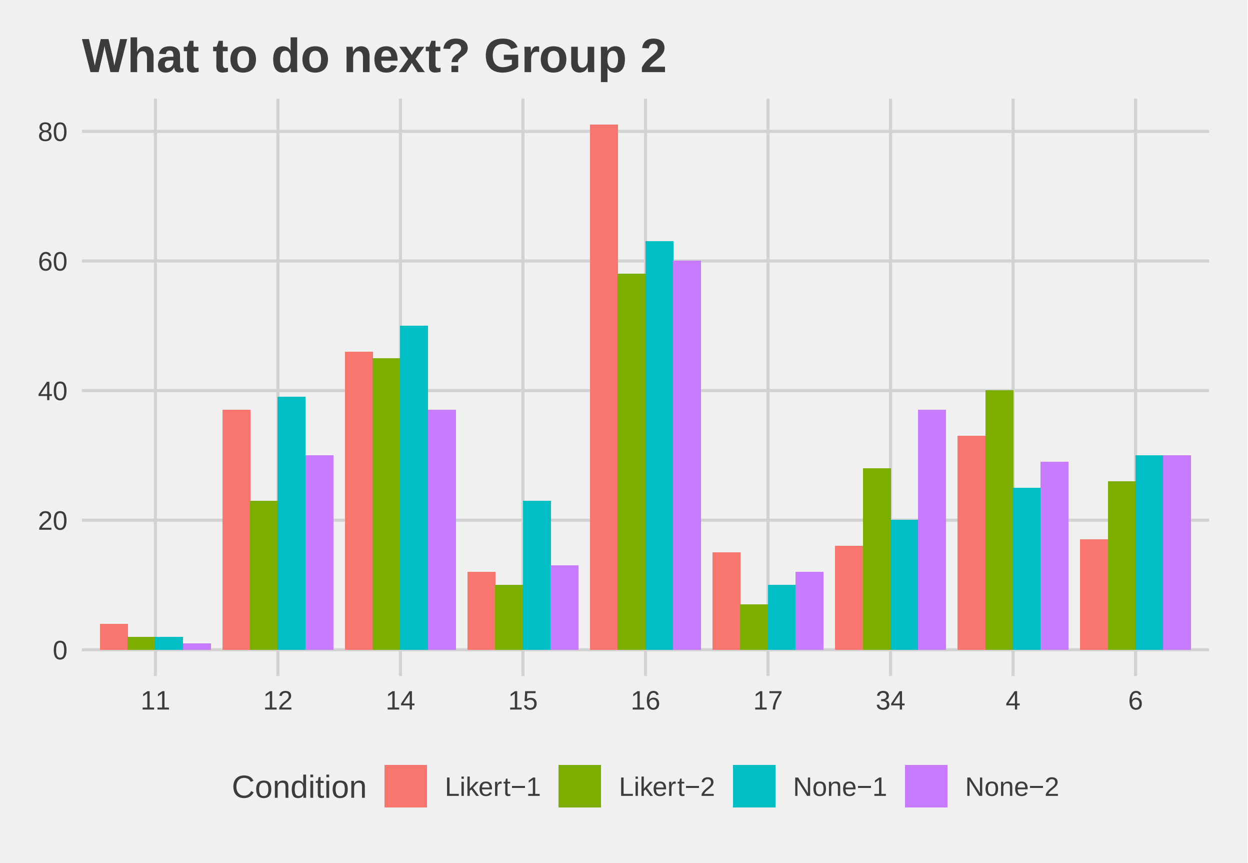
I also constructed three aggregate models for the 11 remaining items. Overall, the effects are pretty small. Putting Group 2 has some small (positive) effect on the fraction of students selecting Group 1 or neither group and, as found above, the Likert items appear to increase the fraction of students that select 'both groups'.

The interaction models indicated, as we found above, that the effects of the Likert items are more variable. I've only shown the 20 largest effects in the last plot.

“What to do next” questions

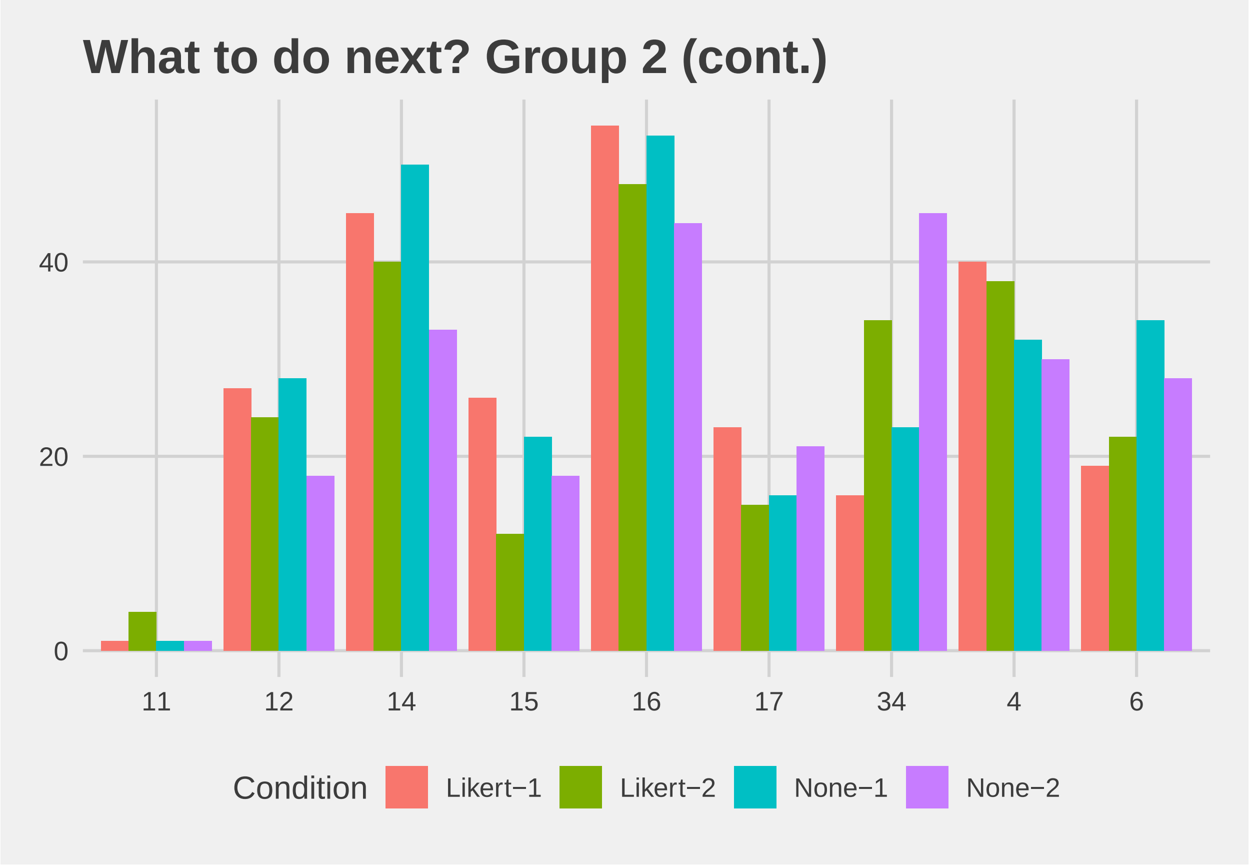


```
##                                                                                 t.info.temp.
## 1                                                           Test or control other variables
## 2  Reduce uncertainty (e.g., more trials for the same masses, more bounces per trial, etc.)
## 10                                                                  Account for human error
## 3                                      Repeat the experiment with more and different masses
## 9                                               Repeat the experiment with better equipment
## 4        Use a different analysis (e.g., graph the results, incorporate systematic effects)
## 6                                              Compare their k-values to the expected value
## 8                                               Design a new experiment to test the results
## 5                                                          Check their work and write it up
## 7                                                                                     Other
## Test for Multiple Marginal Independence (MMI)
##
## Unadjusted Pearson Chi-Square Tests for Independence:
## X^2_S = 38.63
## X^2_S.ij =
##    1   10   2     3    4    5   6    7   8    9
##  3.9 4.52 4.4 10.41 5.31 3.27 2.1 0.99 2.2 1.55
##
## Bootstrap Results:
## 39 resamples were removed from the analysis due to not having all rows or columns represented in an
## Final results based on 1961 resamples
## p.boot = 0.105
## p.combo.prod = 0.1168
## p.combo.min = 0.1188
##
## Second-Order Rao-Scott Adjusted Results:
## X^2_S.adj = 33.09
## df.adj = 25.7
## p.adj = 0.1498
##
## Bonferroni Adjusted Results:
## p.adj = 0.1541
## p.ij.adj =
##  1      10     2      3      4      5      6      7      8      9
##  1.0000 1.0000 1.0000 0.1541 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
```



```
##                                                                      t.info.temp.
## 6                                                     Test or control other variables
## 4                                               Repeat the experiment with better equipment
## 12     Change the analysis (e.g., use a different fit line, incorporate systematic effects)
## 14                                              Compare their k-value to the expected value
## 15                                              Design a new experiment to test the results
## 16 Reduce uncertainty (e.g., more trials for the same masses, more bounces per trial, etc.)
## 17                                                         Check their work and write it up
## 34                                          Repeat the experiment with more and different masses
## 11                                                                    Other (Please describe)
## Test for Multiple Marginal Independence (MMI)
##
## Unadjusted Pearson Chi-Square Tests for Independence:
```

```
## X^2_S = 51.77
## X^2_S.ij =
##    11   12   14   15   16   17    34    4    6
##  1.86 4.05 1.63 6.49 5.82 2.86 15.98 6.64 6.44
##
## Bootstrap Results:
## Final results based on 2000 resamples
## p.boot = 0.0055
## p.combo.prod = 0.0055
## p.combo.min = 0.012
##
## Second-Order Rao-Scott Adjusted Results:
## X^2_S.adj = 45.37
## df.adj = 23.66
## p.adj = 0.0046
##
## Bonferroni Adjusted Results:
## p.adj = 0.0103
## p.ij.adj =
##  11     12     14     15     16     17     34     4      6
##  1.0000 1.0000 1.0000 0.8104 1.0000 1.0000 0.0103 0.7572 0.8298
```



```
##                                                                                t.info.temp.
## 6                                                           Test or control other variables
## 4                                               Repeat the experiment with better equipment
## 12     Change the analysis (e.g., use a different fit line, incorporate systematic effects)
## 14                                              Compare their k-value to the expected value
## 15                                              Design a new experiment to test the results
## 16 Reduce uncertainty (e.g., more trials for the same masses, more bounces per trial, etc.)
## 17                                                         Check their work and write it up
## 34                                     Repeat the experiment with more and different masses
## 11                                                                  Other (Please describe)
## Test for Multiple Marginal Independence (MMI)
## 
## Unadjusted Pearson Chi-Square Tests for Independence:
## X^2_S = 51.22
## X^2_S.ij = 
##    11   12   14   15   16  17    34    4    6
##  4.32 1.88 3.01 4.68 0.55 2.6 25.68 2.48 6.02
## 
## Bootstrap Results:
## 1 resamples were removed from the analysis due to not having all rows or columns represented in an r
## Final results based on 1999 resamples
## p.boot = 0.0065
## p.combo.prod = 0.0045
## p.combo.min < 0.0005
## 
## Second-Order Rao-Scott Adjusted Results:
## X^2_S.adj = 43.59
## df.adj = 22.98
## p.adj = 0.0058
## 
## Bonferroni Adjusted Results:
## p.adj = 0.0001
## p.ij.adj = 
##  11     12     14     15     16     17     34     4      6     
##  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0.0001 1.0000 0.9963
```

MI.test uses three methods for conducting chi-squared tests of independence with multiple response categorical variables (MRCV), which violate regular chi-squared test assumptions of mutual exclusivity. All three methods produced similar p-values for each of the three methods questions (Group 1: p = (0.10, 0.16), Group 2: p = (0.004, 0.011), Group 2, cont: p = (<0.0005, 0.007)), suggesting that we can't say there are any differences in how students respond to Group 1's "what's next?" questions regardless of condition. For Group 2 and Group 2 (cont.), we can make this conclusion, but looking at the individual item chi-squared values and the plots, the differences in distributions are driven by one item in both cases: 34 – "Repeat the experiment with more and different masses" and, again, is mainly affected by the ordering of the groups. In both conditions where students saw Group 2 first, they were more likely to say that the group should test more masses, which makes sense. When students see Group 1 first, then Group 2, the number of masses tested by Group 2 doesn't sound so bad.

Overall, just looking at the plots, the group ordering effect is more pronounced than the effect of the Likert items. Effects of group ordering would look like ||// or //|| (e.g., up/down/up/down), whereas effects of Likert items would look like |||___ or ___||| (e.g., up/up/down/down).