Plausible Alien Zoo: Summary of evaluation and results (October 2021)

Contents

To Dos when real data arrives

- include the 'remove problematic users section' (taken out now, because wouldn't run otherwise with simulated toy data, as too many instances would be removed)

Introduction

This is an analysis of data acquired in the plausible Alien Zoo study run on Amazon mechanical turk in October-November 2021. In this study, naive users were asked to interact with the Alien Zoo paradigm to understand relationships in an unknown dataset, what has been termed "learning to discover" by (Adadi and Berrada 2018). In regular intervals, participants receive counterfactual explanations (CFEs) regarding past choices. These are either "closest" CFEs that fulfill the "smallest feature change" condition (Wachter, Mittelstadt, and Russell 2017), or "plausible" CFEs that are smallest feature changes and also prototypical instances of the data (Artelt and Hammer 2020).

Hypotheses

The main hypothesis is the following:

H1) Plausible CFEs will be more helpful to users tasked to discover unknown relationships in data than closest ones. This should affect objective as well as subjective understandability.

That means, we expect users in the plausible condition to

H1.1) perform better over time in terms of number of Shubs generated, *AND*

H1.2) will become quicker in the final blocks, because choosing the right plants will become more automatic, *AND*

H1.3) can more clearly state which plants were crucial for the Shubs to prosper (questionnaire items 1 and 2)

Further, we expect:

H2) Users will differ in terms of their subjective understanding, specifically:

H2.1) Users will differ in how far they found the explanations useful, and in how far they could made use of it, with an advantage of plausible CFEs (questionnaire items 5, 6)

H2.2) Users imagine plausible CFEs to be more helpful for others users, too (questionnaire item 8).

However:

H3) We do not expect users in different conditions to differ in terms of how well they understood the explanations per se, or needing support for understanding, because explanations are basically the same structurally (questionnaire items 3, 4). So this is also control to make sure groups don't differ in a weird way.

Last:

H4) We expect timing and efficacy of how CFEs were presented to be comparable, as it was literally the same (questionnaire item 9) - a further control.

Finally, we *do not formulate a prediction whether users will uncover inconsistencies in the feedback* (maybe that happens in case of "closest" CFEs when we're in the areas of "no training data"?) (questionnaire item 7). This will be investiagted in a further exploratory analysis.

Descriptive stats

Let's first just look at the data we have.

```
## File .here already exists in /Users/ukuhl/sciebo/IntepretML/Studies/AlienZoo_v01/GitLab/alienzoo/Sta
```

```
##  [1] "userId"    "group"     "blockNo"   "trialNo"   "plant1"    "plant2"
##  [7] "plant3"    "plant4"    "plant5"    "CFplant1"  "CFplant2"  "CFplant3"
## [13] "CFplant4"  "CFplant5"  "shubNoOld" "shubNoNew"
```

How many users do we have in our performance df? 100

Do we have an equal number of users in each dataframe? TRUE

IF NOT, CHECK WHY!

Quality criteria

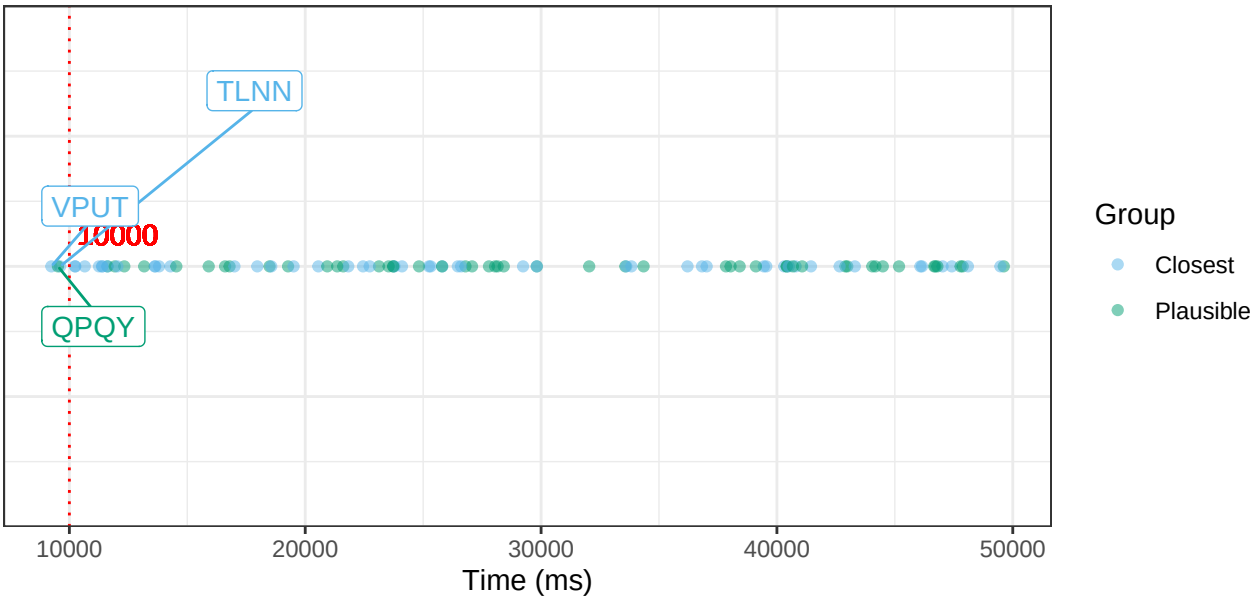
Identify "speeders"

Speeders are people clicking through the study way too quickly (i.e. only a few seconds).

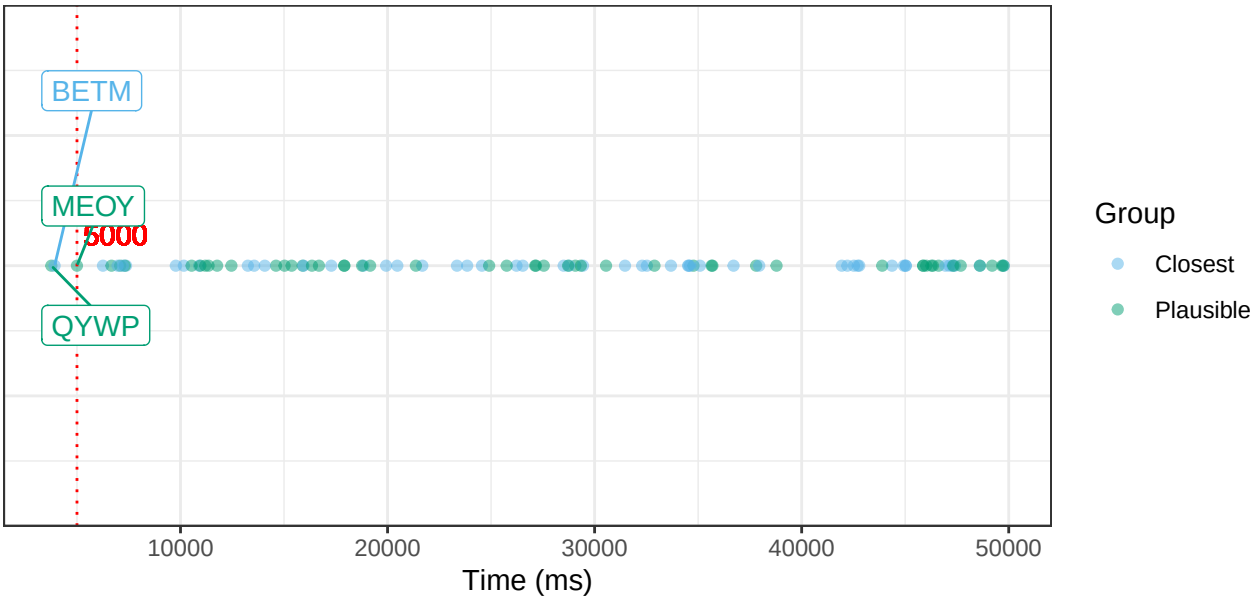
Aim: identify IDs being faster than specified values (variable per game part).

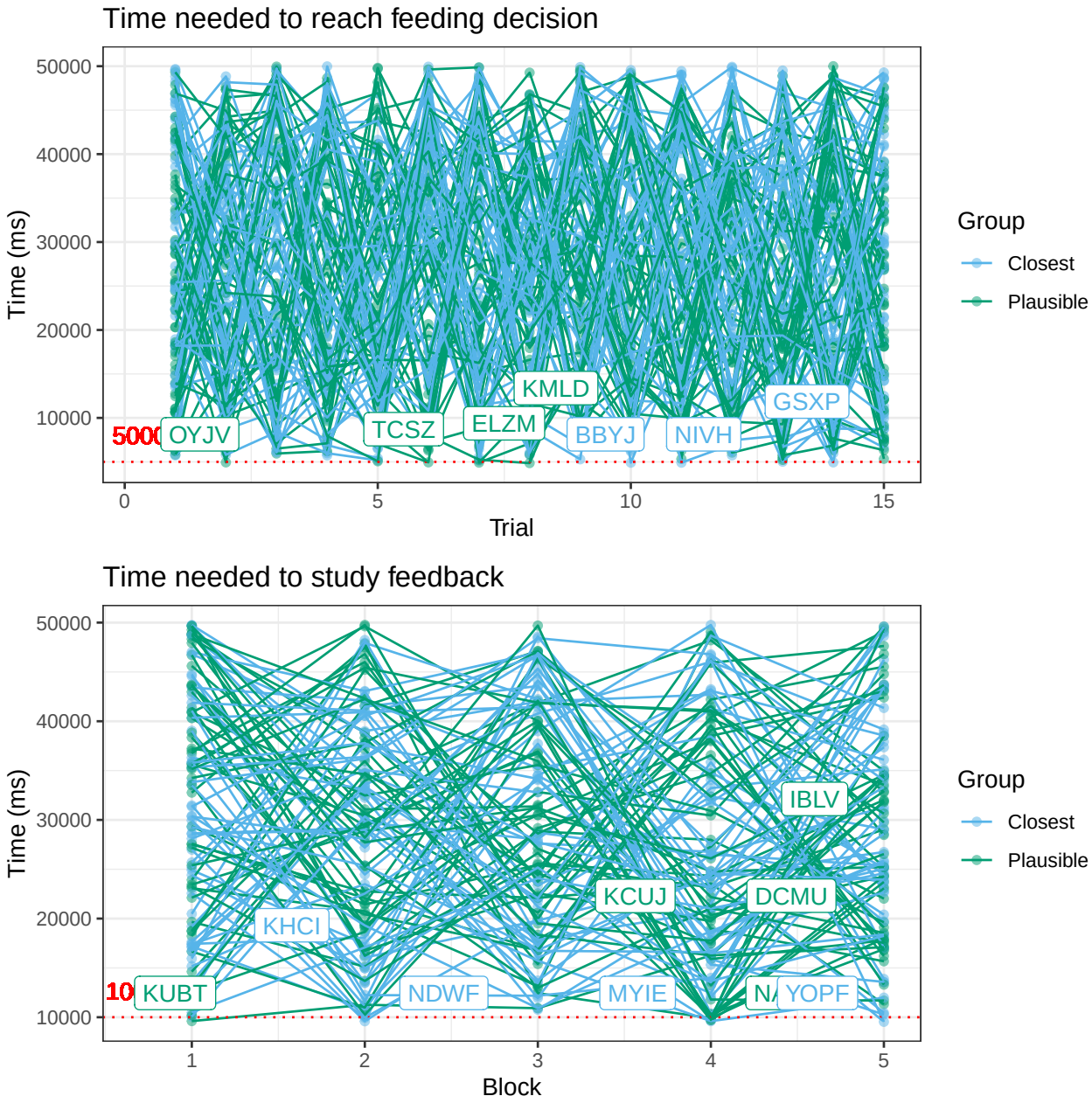
```
## [1] "Display detailed RT data for different trials:"
```

Time spent on agreement scene



Time spent on start (instruction) scene





Identify participants failing the attention check

We include 2 attention checks by asking participants to indicate current pack size after trials 5 and 11.

Aim: Identify IDs of users getting either one or both checks wrong.

Identify "straight-liners" in game part

Identify users who always give the same answer in the game part (over individual blocks, and over all blocks)?

Aim: identify IDs of users "straight-lining" in at least two blocks.

Identify "straight-liners" in survey part

Identify users who always give very uniform answers in the survey part.

Aim: identify IDs of users "straight-lining," i.e. giving only 1 or 2 distinct responses throughout all questions.

Remove data from problematic users

As we have identified users that seem to have dodgy data, we want to remove them.

So to summarize:

- we have 100 users to begin with
- we remove 22 speeders
- we remove 1 users that failed both attention tests
- we remove 0 users that straightlined in the game
- we remove 2 users that straightlined in the survey

Finally: How many users do we have in our clean performance df? 75

Do we have an equal number of users in each clean dataframe? TRUE

IF NOT, CHECK WHY!

Statistical assessment

[...] Comparisons of performance over time between users in the plausible and closest conditions, respectively, are performed using R–4.1.1 (*REF REF REF R Core Team, 2013*). Changes in performance over the XX trials / XX blocks as a measure of learning rate per group are modeled using the lme4 package v.4_1.1-27.1.

The dependent variable is number of Shubs generated. The final model includes the fixed effects of group, trial number and their interaction. The random-effect structure includes a by-subjects random intercept. Advantages of using this approach include that these models account for correlations of data drawn from the same participant (Detry and Ma 2016), account for missing data, and better availability of post-hoc tests (accroding to here: https://www.theanalysisfactor.com/advantages-of-repeated-measures-anova-as-a-mixed-model/; find sciency references!)

Model fits are compared with the analysis of variance function of the stats package. Effect sizes are computed in terms of η_p^2 using the effectsize package v.0.5.

Significant main effects or interactions are followed up by computing the pairwise estimated marginal means. All post-hoc analyses reported are bonerroni corrected to account for multiple comparisons.

%%%%%%%%%%%

Alternative version, if me also account for the slope (try in model, does not seem to work on simulated data). [...] Comparisons of performance over time between users in the plausible and closest conditions, respectively, were performed using R–4.1.1 (*REF REF REF R Core Team, 2013*). Changes in reaction times over the XX trials / XX blocks as a measure of learning rate per group and were modeled using the lme4 package v.4_1.1-27.1, using a maximal random effects structure (*REF REF REF Barr, Levy, Scheepers, & Tily, 2013*). The dependent variable is number of Shubs generated. The final model includes the fixed effects of group, trial number and their interaction. The random-effect structure includes a by-subjects random intercept and a slope for trials. [...]

H1: Plausible CFEs are more helpful to users than closest CFEs

Recap the full hypothesis:

H1) Plausible CFEs will be more helpful to users tasked to discover unknown relationships in data than closest ones. This should affect objective as well as subjective understandability.

That means, we expect users in the plausible condition to

H1.1) perform better over time in terms of number of Shubs generated, *AND*

H1.2) will become quicker in the final blocks, because choosing the right plants will become more automatic, *AND*

H1.3) can more clearly state which plants were crucial for the Shubs to prosper (questionnaire items 1 and 2)

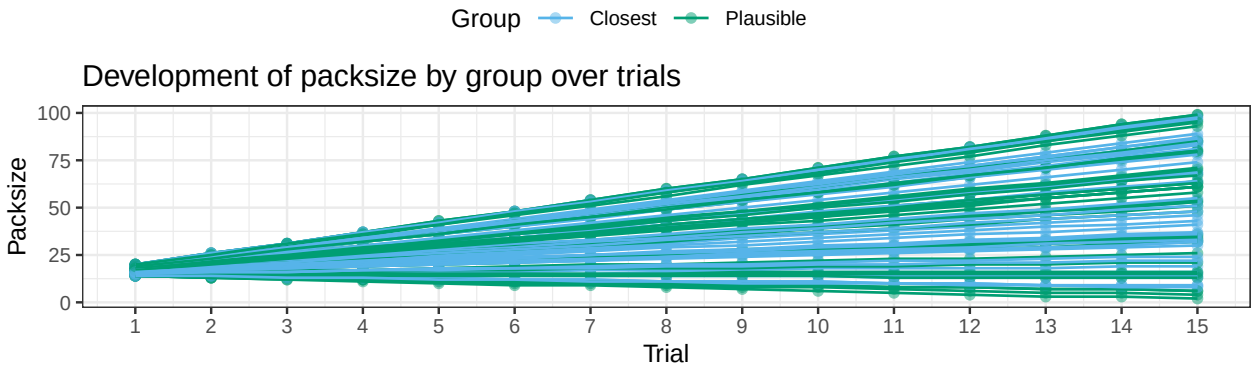
H1.1) Users in the plausible condition perform better over time in terms of number of Shubs generated

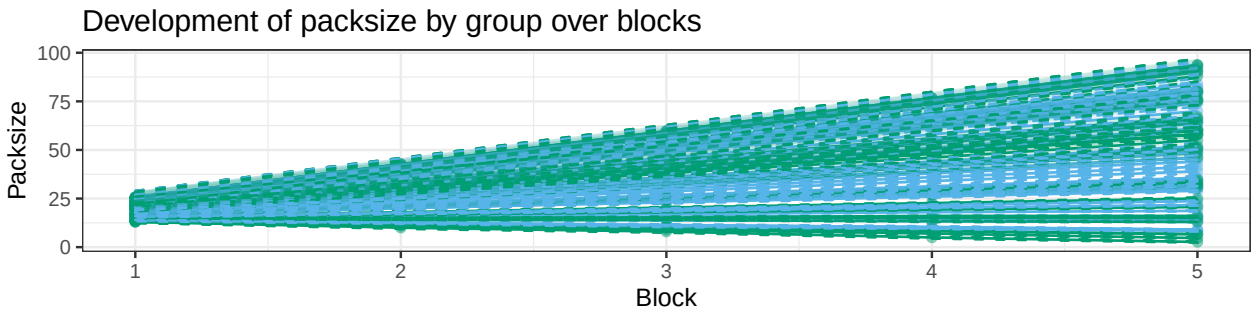
Let's start with a first peek at the data: Descriptive stats + plotting the packsize trajectories per trial and block for each person individually.

```
## [1] "First peek at the data, getting min / max / median:"
```

```
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   18.00   28.00   33.92   45.50   98.00
##
## $P
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   18.00   30.00   35.07   48.00   99.00
```

```
## [1] "Display figures showing development of packsize over trials / blocks:"
```





Now on to the statistics.

```
## [1] "ANOVA table:"
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                 SumSq MeanSq NumDF DenDF  Fvalue  Pvalue
## group               8    7.8     1    73   0.103 0.74869
## trialNo        130995 9356.8    14  1022 124.546 0.00000
## group:trialNo     106    7.5    14  1022   0.100 0.99999
```

```
## NOTE: Results may be misleading due to involvement in interactions
## NOTE: Results may be misleading due to involvement in interactions
```

Results The analysis revealed:

- main effect of group: F(1,72.9999971)=0.1034167, p=0.7486864,η_p^2=0.0014147
- main effect of trials (time): F(14,1022.0000014)=124.5464924, p=0,η_p^2=0.6304667
- interaction (group x trials): F(14,1022.0000014)=0.1004637, p=0.9999905,η_p^2=0.0013743

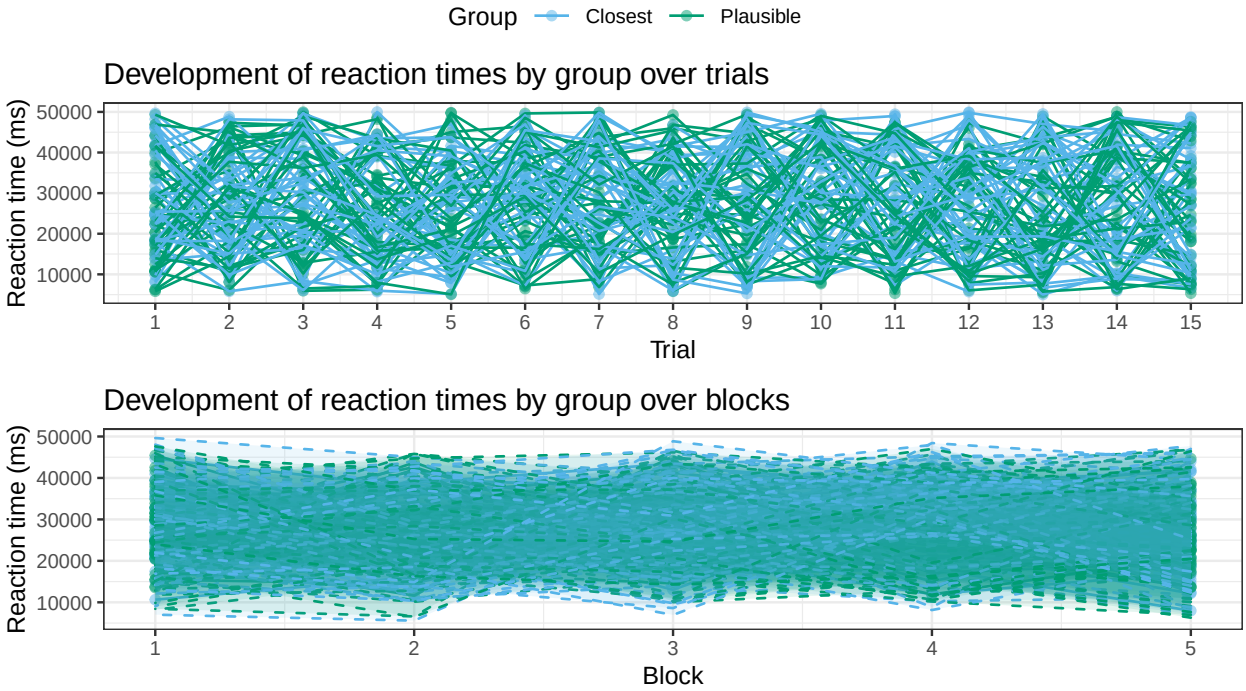
H1.2) Users in the plausible condition become quicker in deciding what plants to choose in the final blocks, because choice of the right plants will become more automatic

Again, first peek at the data: Descriptive stats + plotting the RT trajectories per trial and block for each person individually.

```
## [1] "First peek at the data, getting min / max / median:"
```

```
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5020   16278   28768   27997   39468   49976
##
## $P
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5066   16102   26562   27244   38914   49999
```

```
## [1] "Display figures showing development of reaction times over trials / blocks:"
```



Now on to the statistics.

```
## [1] "ANOVA table:"
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                     SumSq    MeanSq NumDF DenDF  Fvalue  Pvalue
## group           118653219 118653219     1    73 0.70806 0.40284
## TrialNr        2097743586 149838828    14  1022 0.89417 0.56501
## group:TrialNr  2119426814 151387630    14  1022 0.90341 0.55475
```

```
## NOTE: Results may be misleading due to involvement in interactions
## NOTE: Results may be misleading due to involvement in interactions
```

The analysis revealed:

- main effect of group: F(1,73.0000001)=0.7080648, p=0.4028351,η^2_p=0.0096063
- main effect of trials (time): F(14,1022.0000002)=0.8941654, p=0.5650051,η^2_p=0.0121006
- interaction (group x trials): F(14,1022.0000003)=0.9034079, p=0.5547495,η^2_p=0.0122242

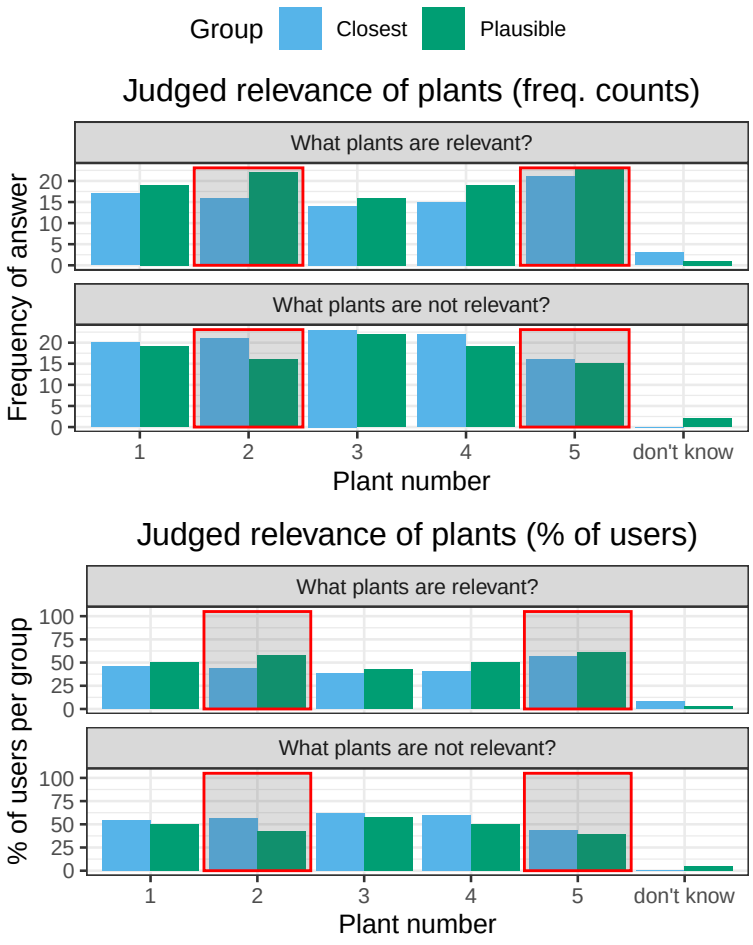
H1.3) Users in the plausible condition can more clearly state which plants were crucial for the Shubs to prosper (questionnaire items 1 and 2)

Questionnaire items 1 and 2 explicitly ask users to state which plants they thought were relevant. So what did users tick?

```
##     userId             group           itemNo  responseNo    checked
##  Length:900         Length:900         1:450   1:150      Min.   :0.0000
##  Class :character   Class :character   2:450   2:150      1st Qu.:0.0000
##  Mode  :character   Mode  :character           3:150      Median :0.0000
##                                                4:150      Mean   :0.4233
##                                                5:150      3rd Qu.:1.0000
##                                                6:150      Max.   :1.0000
```

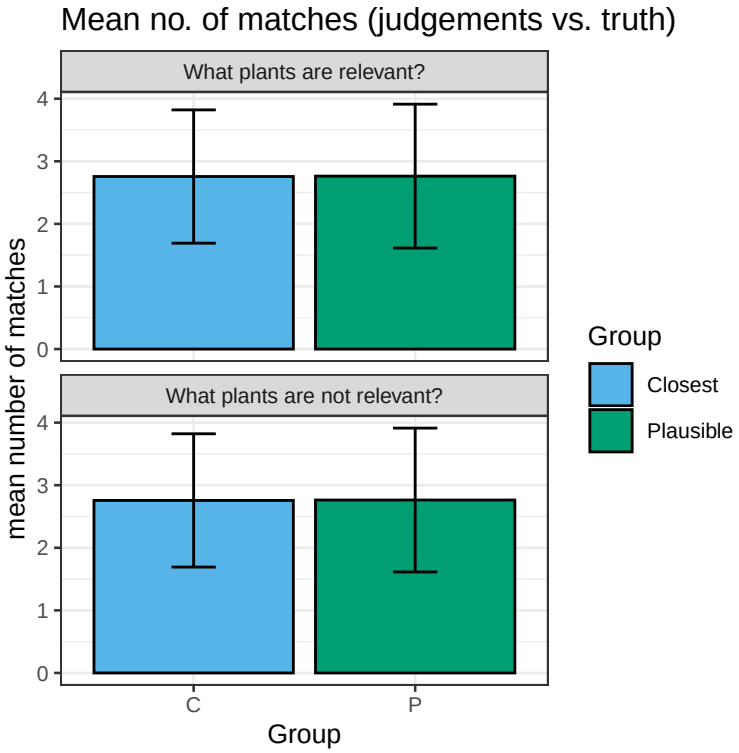
```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

```
## [1] "Display figures showing user responses in relevant survey items:"
```



How to evaluate this statistically? Let's just count the matches between 'judged as relevant' / 'judged as irrelevant' user vectors and the true 'relevant' / 'irrelevant' factors.

```
## [1] "Mean number of matches between user judgements and ground truth for relevant and irrelevant pla
```



The analysis revealed:

- Is there a significant difference in terms of matches between plants judged as relevant and ground truth?: We compared number of matches for users in plausible condition (M = 2.7631579, SD = 1.1492468) and users in the closest condition (M = 2.7567568, SD = 1.0647223) using a Wilcox test. This showed
  - for ttest: t()=708, p=0.9605764, cohen's d = 0.0057077
  - for wilcoxon test: U=708, p=0.9605764, r = 0.0057077
- Is there a significant difference in terms of matches between plants judged as irrelevant and ground truth?: We compared number of matches for users in plausible condition (M = 2.7631579, SD = 1.1492468) and users in the closest condition (M = 2.7567568, SD = 1.0647223) using a Wilcox test. This showed
  - for ttest: t()=708, p=0.9605764, cohen's d = 0.0057077
  - for wilcoxon test: U=708, p=0.9605764, r = 0.0057077

H2) User differences in terms of subjective understanding

Recap the full hypothesis:

H2) Users will differ in terms of their subjective understanding, specifically:

H2.1) Users will differ in how far they found the explanations useful, and in how far they could made use of it, with an advantage of plausible CFEs (questionnaire items 5, 6)

H2.2) Users imagine plausible CFEs to be more helpful for others users, too (questionnaire item 8).

H2.1) Users will differ in how far they found the explanations useful, and in how far they could made use of it, with an advantage of plausible CFEs (questionnaire items 5, 6)

Diving more into survey results.

Item 5: "I found that the feedback on what choice would have led to a better result helped me to increase the number of Shubs."

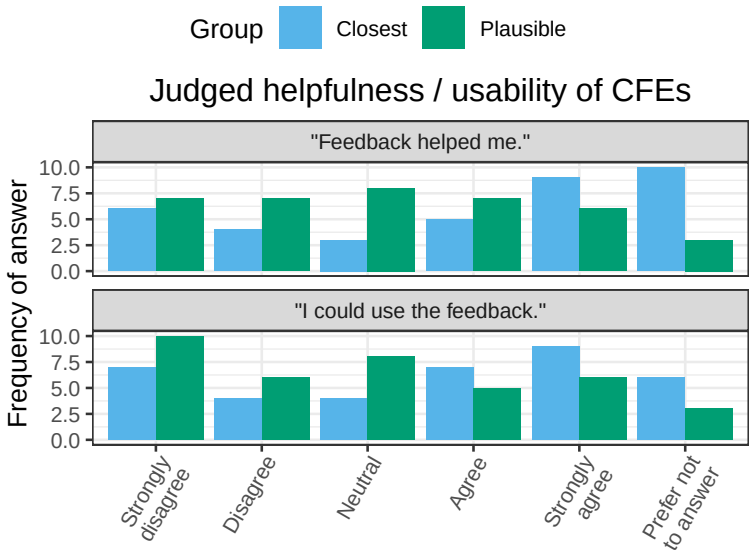
Item 6: "I was able to use the feedback based on what choice would have led to a better result to increase the number of Shubs."

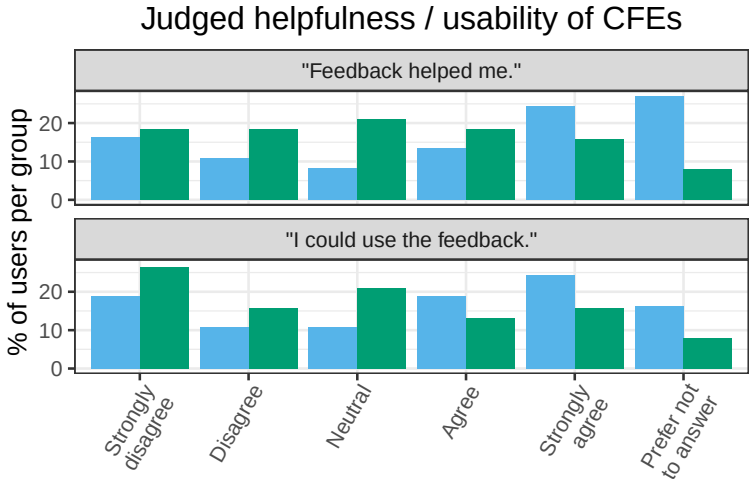
We will talk about these as quantifying how subjectively helpful (item 5) and how usable (item 6) they were.

```
##     userId             group           itemNo  responseNo    checked
##  Length:900         Length:900         5:450   1:150      Min.   :0.0000
##  Class :character   Class :character   6:450   2:150      1st Qu.:0.0000
##  Mode  :character   Mode  :character           3:150      Median :0.0000
##                                                4:150      Mean   :0.1667
##                                                5:150      3rd Qu.:0.0000
##                                                6:150      Max.   :1.0000
```

```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

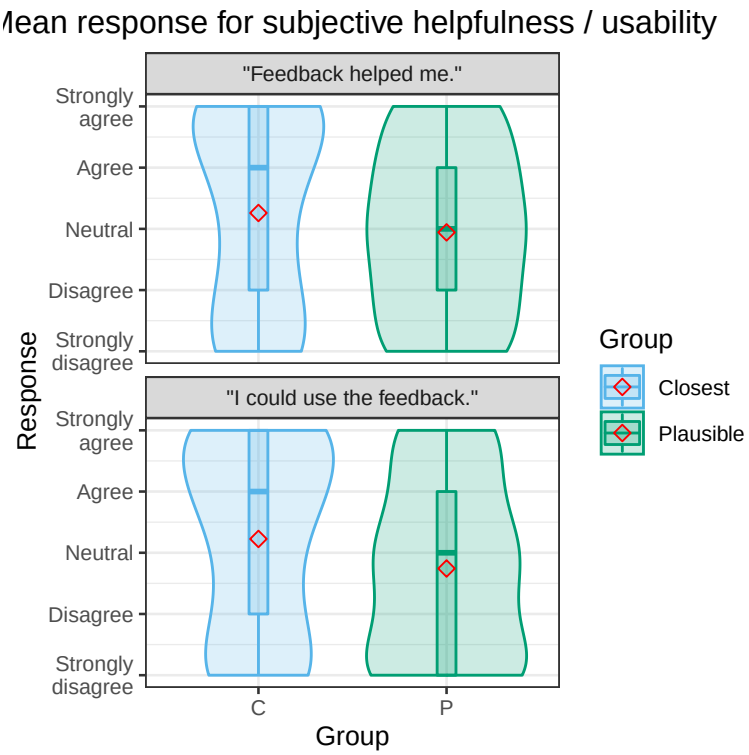
```
## [1] "Display figures showing user responses in relevant survey items:"
```





On to the statistical comparison: for Likert-scale, we want a non-parametric statistical test for ordinal data, that's the Wilcoxon–Mann–Whitney U test.

```
## [1] "Mean user response for subjective helpfulness / usability:"
```



The analysis revealed:

- Is there a significant difference in terms of subjective helpfulness between groups? We compared responses for subjective helpfulness for users in plausible condition (M = 2.9428571, SD = 1.3920543) and users in the closest condition (M = 3.2592593, SD = 1.6074967) using a Wilcoxon–Mann–Whitney U test. This showed: U=412.5, p=0.3882184, r = -0.1095824
- Is there a significant difference in terms of subjective usability?: We compared responses for subjective usability for users in plausible condition (M = 2.7428571, SD = 1.4621269) and users in the closest condition (M = 3.2258065, SD = 1.5643886) using a Wilcoxon–Mann–Whitney U test. This showed: U=444.5, p=0.200101, r = -0.1577127

H2.2) Users imagine plausible CFEs to be more helpful for others users, too (questionnaire item 8).

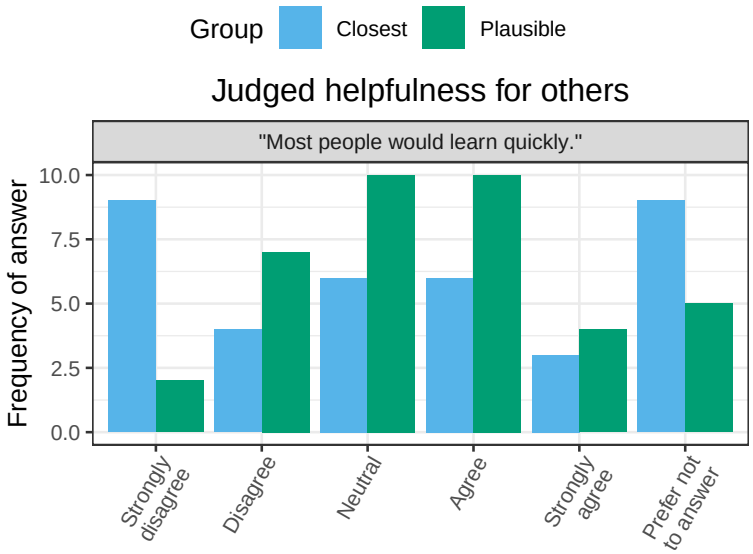
Item 8: "I think most people would learn to work with the feedback on what choice would have led to a better result very quickly."

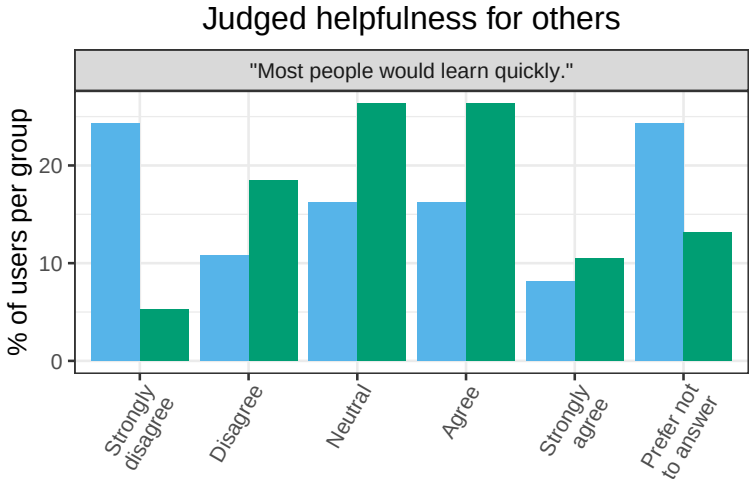
Do users in the plausible condition imagine that explanations would be more helpful for other users, compared to users in the closes condition?

```
##     userId             group               itemNo  responseNo    checked
##  Length:450         Length:450         Min.   :8   1:75       Min.   :0.0000
##  Class :character   Class :character   1st Qu.:8   2:75       1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :8   3:75       Median :0.0000
##                                        Mean   :8   4:75       Mean   :0.1667
##                                        3rd Qu.:8   5:75       3rd Qu.:0.0000
##                                        Max.   :8   6:75       Max.   :1.0000
```

```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

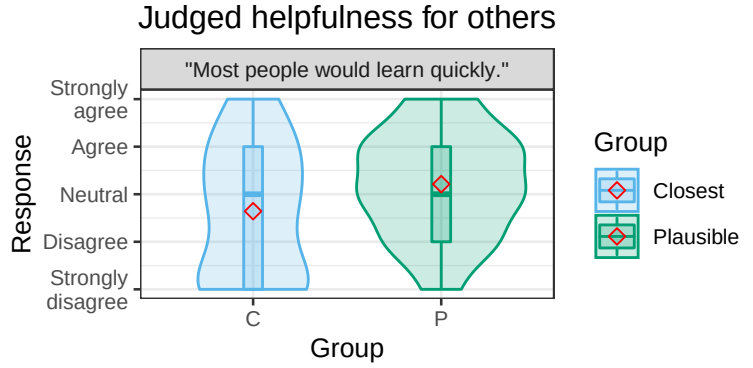
```
## [1] "Display figures showing user responses in relevant survey items:"
```





Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

```
## [1] "Mean user response for subjective helpfulness / usability:"
```



The analysis revealed:

- Is there a significant difference in terms of estimated usefulness for others between groups? We compared number of matches for users in plausible condition (M = 3.2121212, SD = 1.1112374) and users in the

closest condition (M = 2.6428571, SD = 1.4198144) using a Wilcoxon–Mann–Whitney U test. This showed: U=572, p=0.1041256, r = 0.2080821

H3) No expected differences in understanding the explanations per se

Coming to areas where we do not expect differences between groups. CAREFUL though: Remember that Null findings cannot be interpreted, so discuss with caution. However, this may act as an important control to make sure groups don't differ in a weird way.

Revisiting the hypothesis:

H3) We do not expect users in different conditions to differ in terms of how well they understood the explanations per se, or needing support for understanding, because explanations are basically the same structurally (questionnaire items 3, 4).

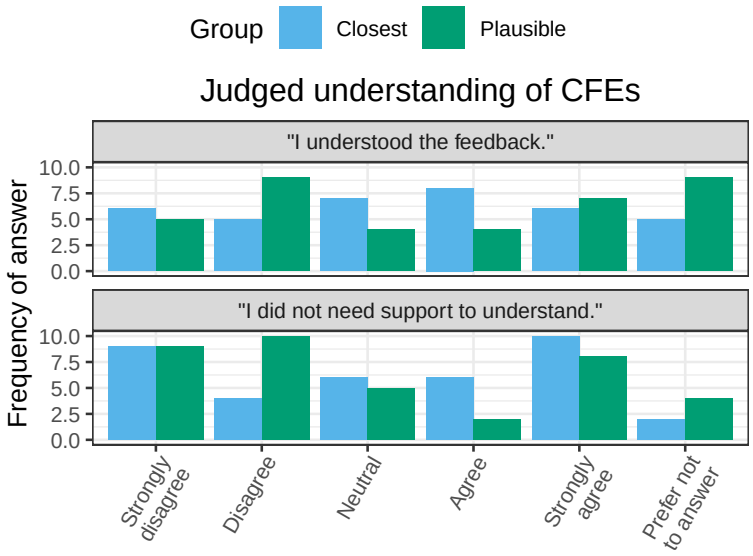
Item 3: "I understood the feedback on what choice would have led to a better result."

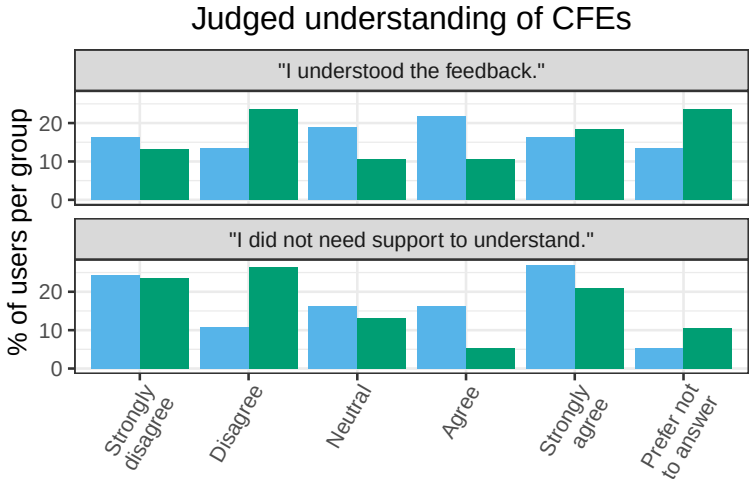
Item 4: "I did not need support to understand the feedback on what choice would have led to a better result."

```
##     userId             group           itemNo  responseNo    checked
##  Length:900         Length:900         3:450   1:150      Min.   :0.0000
##  Class :character   Class :character   4:450   2:150      1st Qu.:0.0000
##  Mode  :character   Mode  :character           3:150      Median :0.0000
##                                                4:150      Mean   :0.1667
##                                                5:150      3rd Qu.:0.0000
##                                                6:150      Max.   :1.0000
```

```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

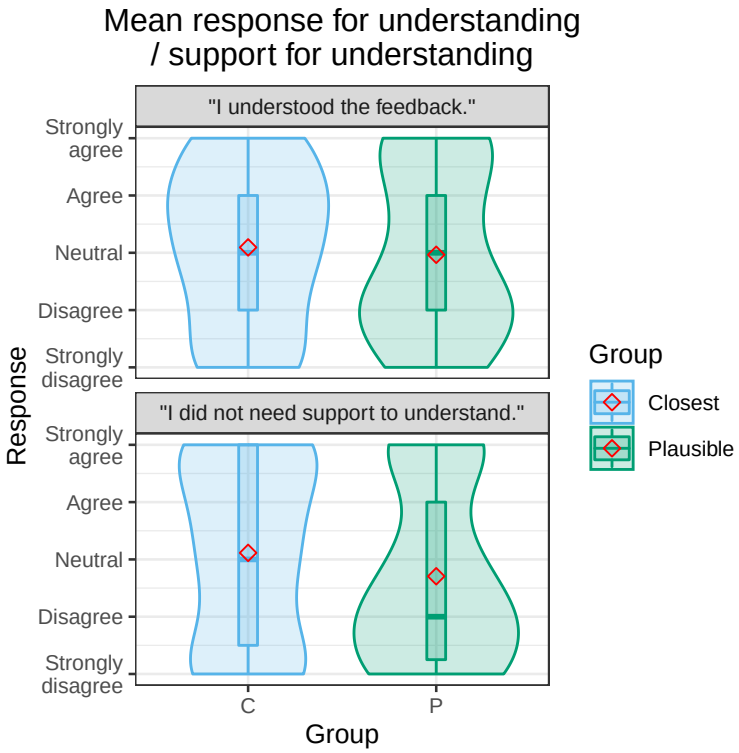
```
## [1] "Display figures showing user responses in relevant survey items:"
```





On to the statistical comparison: for Likert-scale, we want a non-parametric statistial test for ordinal data, that's the Wilcoxon–Mann–Whitney U test.

```
## [1] "Mean user response for understanding / need for support to understand:"
```



The analysis revealed:

- Is there a significant difference in terms of understanding of explanations between groups? We compared responses of users in plausible condition (M = 2.9655172, SD = 1.4755812) and users in the closest condition (M = 3.09375, SD = 1.3995247) using a Wilcoxon–Mann–Whitney U test. This showed: U=457.5, p=0.9293373, r = -0.0113541
- Is there a significant difference in terms of needing support to understand explanations?: We compared responses of users in plausible condition (M = 2.7058824, SD = 1.5281086) and users in the closest condition (M = 3.1142857, SD = 1.5861798) using a Wilcoxon–Mann–Whitney U test. This showed: U=514.5, p=0.3247431, r = -0.1185509

H4) Presented timing and efficacy of how CFEs were presented expected to be comparable

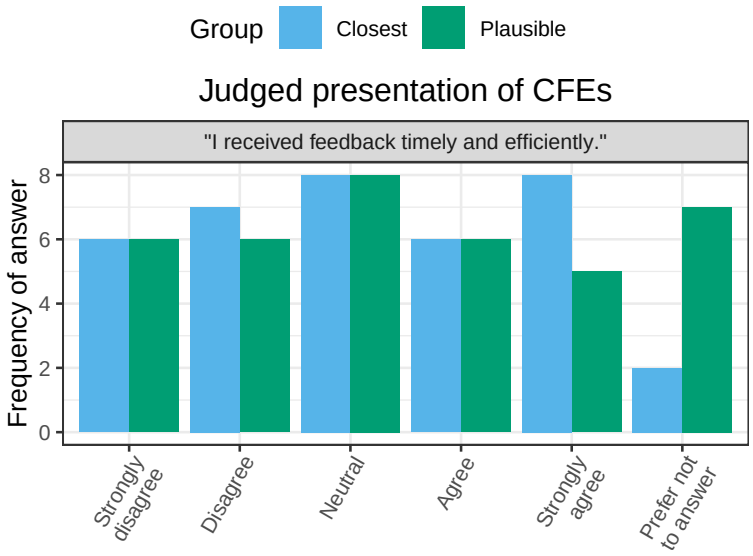
H4) We expect timing and efficacy of how CFEs were presented to be comparable, as it was literally the same (questionnaire item 9) - a further control.

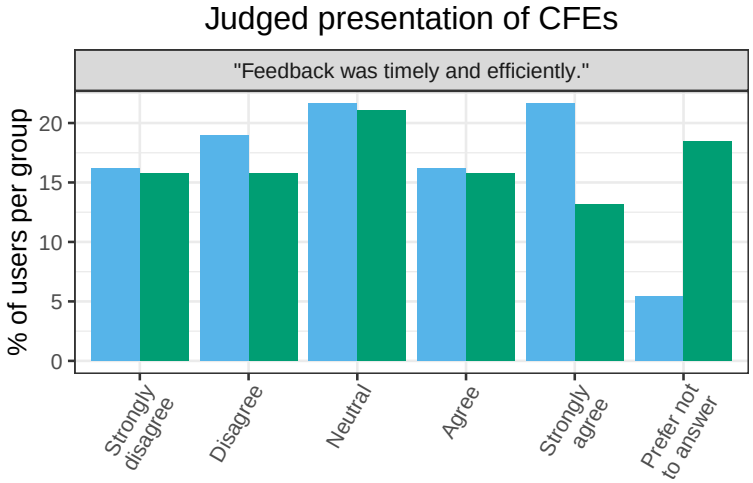
Item 9: "I received the feedback on what choice would have led to a better result in a timely and efficient manner."

```
##     userId             group               itemNo  responseNo    checked
##  Length:450         Length:450         Min.   :9   1:75       Min.   :0.0000
##  Class :character   Class :character   1st Qu.:9   2:75       1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :9   3:75       Median :0.0000
##                                        Mean   :9   4:75       Mean   :0.1667
##                                        3rd Qu.:9   5:75       3rd Qu.:0.0000
##                                        Max.   :9   6:75       Max.   :1.0000
```

```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

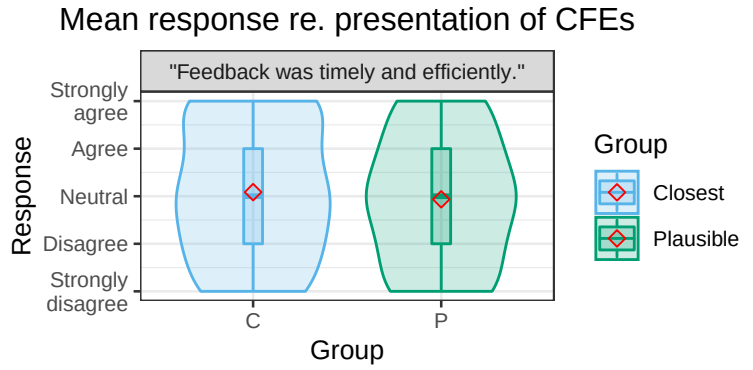
```
## [1] "Display figures showing user responses in relevant survey items:"
```





Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

```
## [1] "Mean user response for subjective helpfulness / usability:"
```



The analysis revealed:

- Is there a significant difference in terms of estimated usefulness for others between groups? We compared number of matches for users in plausible condition (M = 2.9354839, SD = 1.3646852) and users in the

closest condition (M = 3.0857143, SD = 1.4219173) using a Wilcoxon–Mann–Whitney U test. This showed: U=510, p=0.6745526, r = -0.0516871

Final exploratory analysis

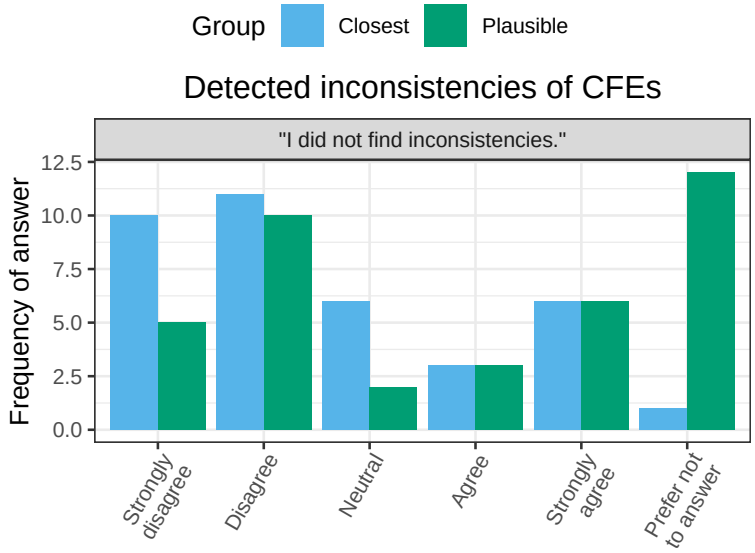
It is not clear whether users were uncover inconsistencies in the feedback. Maybe that is the case for "closest" CFEs when we're in the areas of "no training data"? Let's see what users responded.

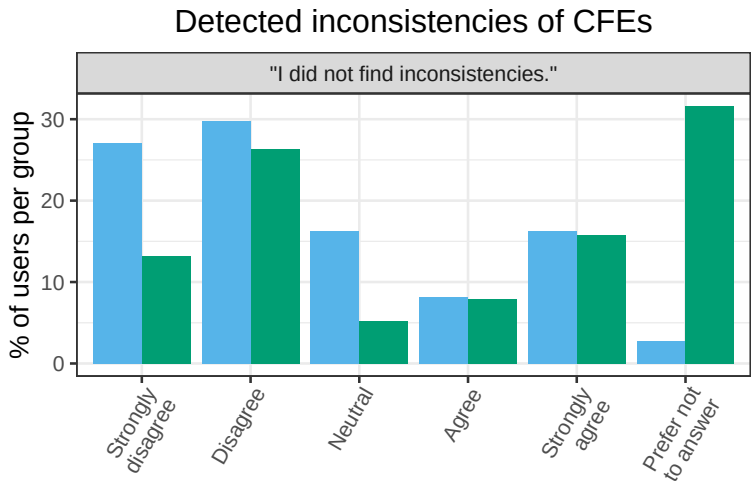
Item 7: "I did not find inconsistencies in the feedback on what choice would have led to a better result."

```
##     userId             group               itemNo  responseNo    checked
##  Length:450         Length:450         Min.   :7   1:75       Min.   :0.0000
##  Class :character   Class :character   1st Qu.:7   2:75       1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :7   3:75       Median :0.0000
##                                        Mean   :7   4:75       Mean   :0.1667
##                                        3rd Qu.:7   5:75       3rd Qu.:0.0000
##                                        Max.   :7   6:75       Max.   :1.0000
```

```
## `summarise()` has grouped output by 'group', 'itemNo'. You can override using the `.groups` argument
```

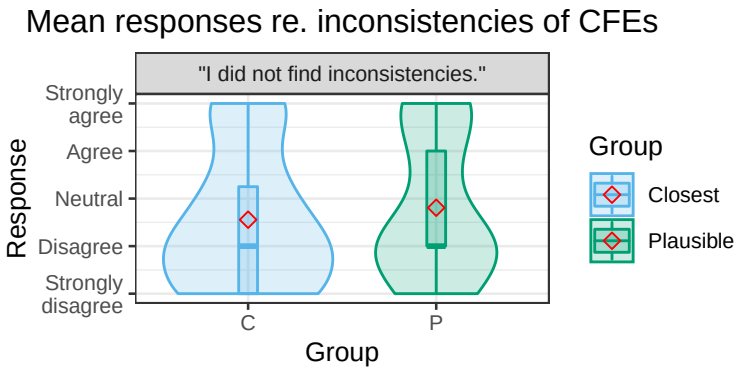
```
## [1] "Display figures showing user responses in relevant survey items:"
```





Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

```
## [1] "Mean user response for inconsistencies of CFEs:"
```



The analysis revealed:

- Is there a significant difference in terms of estimated usefulness for others between groups? We compared number of matches for users in plausible condition (M = 2.8076923, SD = 1.4971768) and users in the closest condition (M = 2.5555556, SD = 1.4231644) using a Wilcoxon–Mann–Whitney U test. This showed: U=511.5, p=0.5262613, r = 0.080483

Wrapping up

```
## [1] TRUE
```

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Artelt, André, and Barbara Hammer. 2020. "Convex Density Constraints for Computing Plausible Counterfactual Explanations." In *Artificial Neural Networks and Machine Learning – ICANN 2020*, edited by Igor Farkaš, Paolo Masulli, and Stefan Wermter, 12396:353–65. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-61609-0_28.

Detry, Michelle A., and Yan Ma. 2016. "Analyzing Repeated Measurements Using Mixed Models." *JAMA* 315 (4): 407. https://doi.org/10.1001/jama.2015.19394.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289.