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

Contents

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).

First things first: rough data cleaning

Let's first just look at the data we have. Excluding all users that had incomplete datasets, what is the turnout?

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

How many users do we have in our performance df before any cleaning (i.e., also including users with incomplete datasets)? 134

How many users do we have in our performance df after cleaning? 101

How many users do we have in our reaction time df after cleaning? 101

How many users do we have in our attention df after cleaning? 101

How many users do we have in our survey df after cleaning? 101

General infos after removal of incomplete datasets

At this point, we have 101 participants.

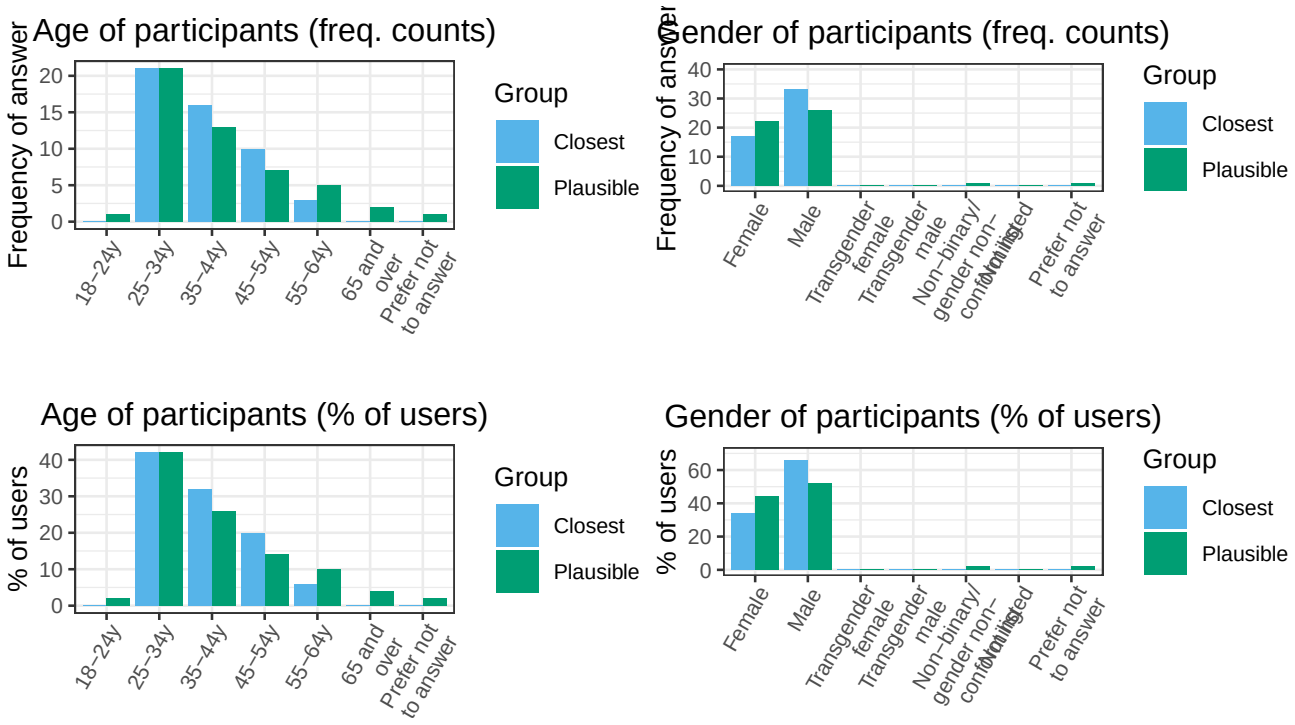
Why 101? We actually acquired 100, but one participant timed out on AMT. That means that he/she played the game and data was fully logged, BUT AMT did not accept their code anymore. Unfortunately, we cannot re-imburse this person for their efforts. Thus, we will exclude that person too.

After removal, check again:

- At this point, we have 100 participants. Of those,
- 50 participants were in the closest condition and
- 50 participants in the plausible condition.

Check covariates across groups

Additionally to assessing performance, we also acquire age and gender information of participants. How do our groups look like? Are the groups comparable?



Let's run a statistical comparison between our two groups. For age, we have ordinal data (in age bands), so we will use a non-parametric statistical test for ordinal data, that's the Wilcoxon–Mann–Whitney U test.

For gender, we need to check if data is normally distributed. If so, use a ttest, if not, we will also use the non-parametric Wilcoxon–Mann–Whitney U.

We acquired data from 100, with 50 users in the control group (17 female, 33 male, median age group is 25-34years), and 50 users in the plausible group (22 female, 26 male, 1 non-binary / gender non-conforming, 1 user did not disclose gender information, median age group is 25-34years).

The analysis showed for *Age*:

- We have age information for 49 users in the plausible and 50 users in the closest group (1 user(s) preferred not to disclose age information).
- Is there a significant difference in terms of age between the groups? We compared number of matches for users in plausible condition and users in the closest condition using a Wilcox test. This showed: U=1234, p=0.9498057, r = 0.0063268

The analysis showed for *Gender*:

- We have gender information for 49 users in the plausible and 50 users in the closest group (1 user(s) preferred not to give gender information).
- Is there a significant difference in terms of gender between the groups? We compared number of matches for users in plausible condition and users in the closest condition using a Wilcox test. This showed
  - for wilcoxon test: U=1108, p=0.3390422, r = -0.0960876

Quality criteria

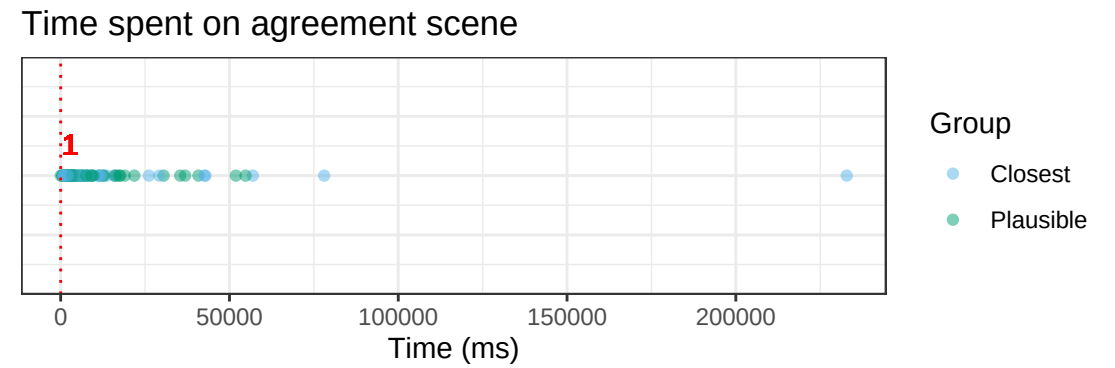
Before going into the hypotheses, we should apply some quality criteria to our data. Sub-quality data should be removed. The following subsections take care of such cases.

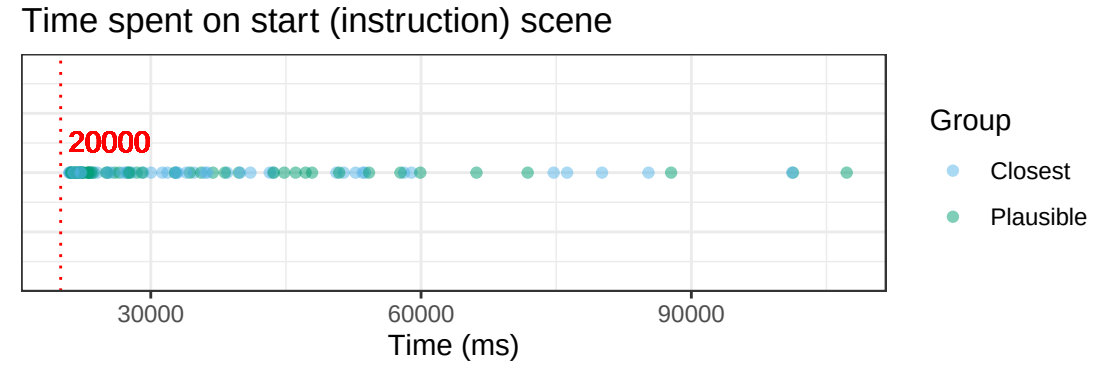
Identify "speeders"

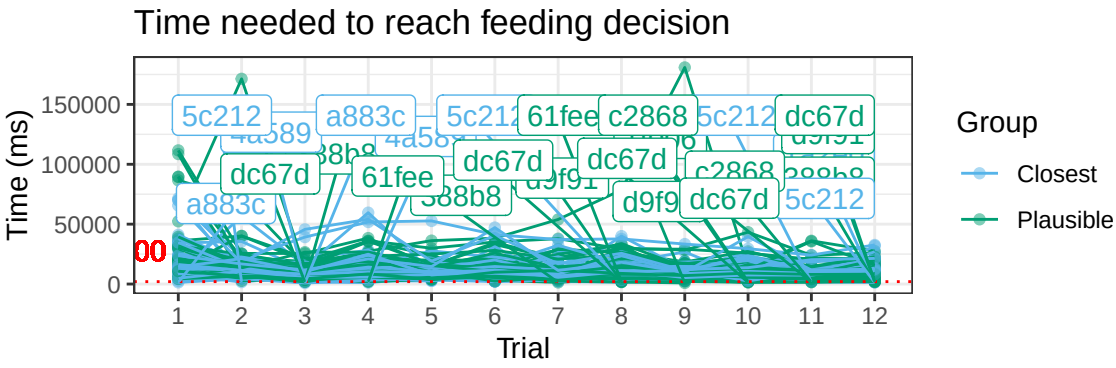
Speeders are people clicking through the study way too quickly to do the task properly.

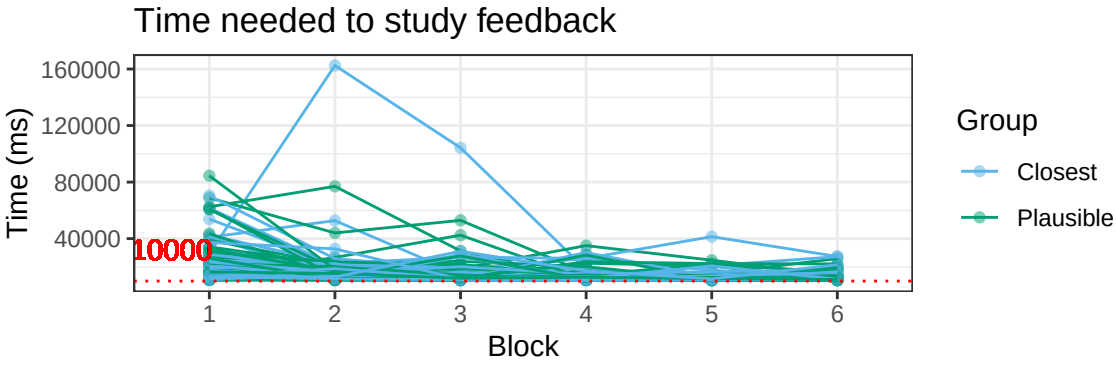
Aim: identify IDs being faster than specified values (variable per game part). This part will tag users that needed less than 2000ms to reach a feeding decision (suspiciously quick) in 4 or more trials.

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









Identify participants failing the two attention checks

We include 2 attention checks during the game by asking participants to indicate current pack size after trials 3 and 7.

Aim: Identify IDs of users getting either one or both checks wrong; exclude those getting both 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) DESPITE not increasing their pack size.

Aim: identify IDs of users "straight-lining" in at least two blocks, while pack size did not change (i.e., who were "immune to feedback").

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 responses with either positive or negative valence.

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 134 users to begin with
- we remove 29 users that have incomplete datasets (aborted prematurely)
- we remove 4 whose information was not logged properly
- 1 participant timed out and could not be re-imbursed properly
- we remove 2 speeders
- we remove 5 users that failed both attention tests during the game
- we remove 3 users that failed the attention test in the survey
- we remove 4 users that straightlined in the game, despite not improving
- remove 12 users that straightlined in the survey

Finally: How many users do we have in our clean performance df? 74

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

To sum up, in our final data we have 74 users, with 40 users in the control group (13 female, 27 male, median age group is 25-34years), and 34 users in the plausible group (18 female, 15 male, 1 non-binary / gender non-conforming, median age group is 35-44years).

Re-check: are there still no significant differences in terms of gender / age?

The analysis showed for *Age* in the clean dataset:

- We have age information for 34 users in the plausible and 40 users in the closest group (0 user(s) preferred not to disclose age information).
- Is there a significant difference in terms of age between the groups? We compared number of matches for users in plausible condition and users in the closest condition using a Wilcox test. This showed: U=712.5, p=0.7178265, r = 0.0420078

The analysis showed for *Gender* in the clean dataset:

- We have gender information for 34 users in the plausible and 40 users in the closest group (0 user(s) preferred not to give gender information).
- Is there a significant difference in terms of gender between the groups? We compared number of matches for users in plausible condition and users in the closest condition using a Wilcox test. This showed

– for wilcoxon test: U=554.5, p=0.1159299, r = -0.1827517

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 9).

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 10) - 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 8). This will be investiagted in a further exploratory analysis.

Statistical assessment

[...] Comparisons of performance over time between users in the plausible and closest conditions, respectively, are performed using R–4.1.1 (R Core Team 2021). Changes in performance over 12 trials as a measure of learning rate per group are modeled using the lme4 package v.4_1.1-27.1.

In the model testing for differences in terms of user performance, the dependent variable is number of Shubs generated. In teh assessment of user's reaction time, we used time needed to reach a feeding decision in each trial as dependent variable. The final models include 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 (according to here: https://www.theanalysisfactor.com/advantages-of-repeated-measures-anova-as-a-mixed-model/; find a more appropriate reference.)

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.

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)

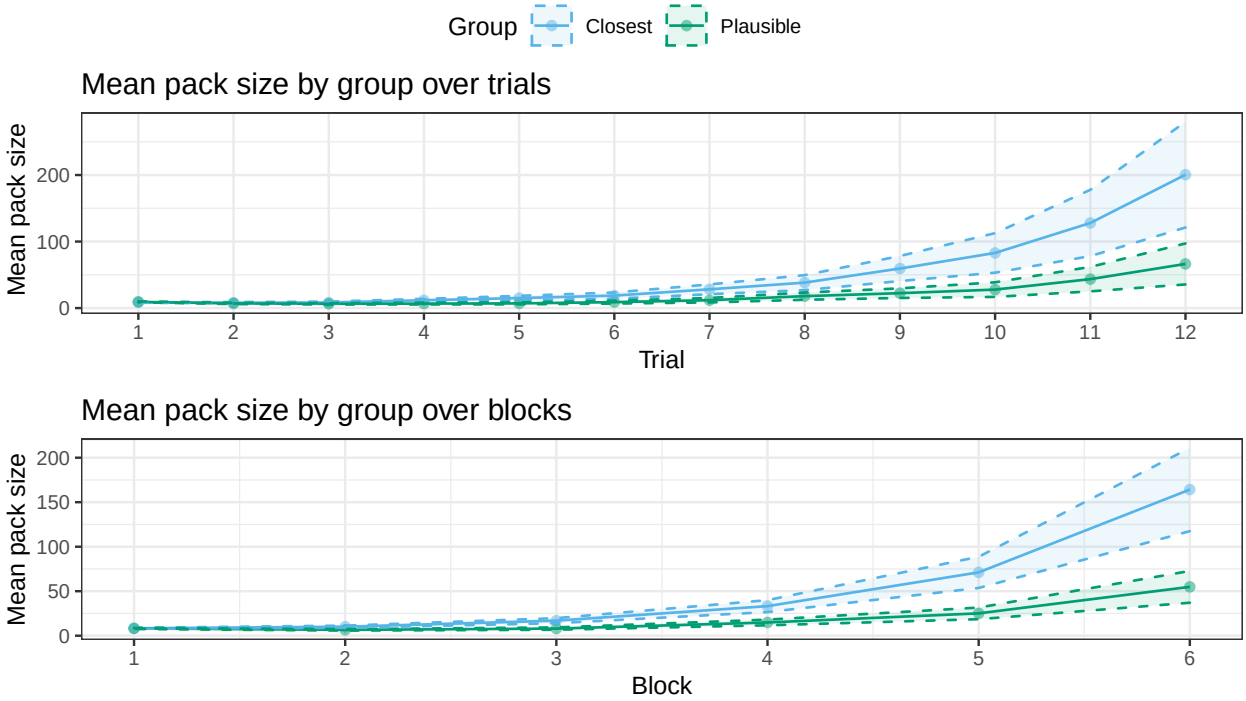
H1.1) Users in the plausible condition perform better over time in terms of number of Shubs generated

Let's start with a first peek at the data: Descriptive stats + plotting the pack size 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.
##    2.00    2.00    6.00   50.65   25.00 2222.00
##
## $P
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.00    3.00   19.59   13.00  754.00
```

```
## [1] "Display figures showing development of pack size 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
## trialNo       1238783 112617    11   792 7.5851 0.000000
## group           38391  38391     1    72 2.5857 0.112207
## trialNo:group  346121  31466    11   792 2.1193 0.017077
```

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

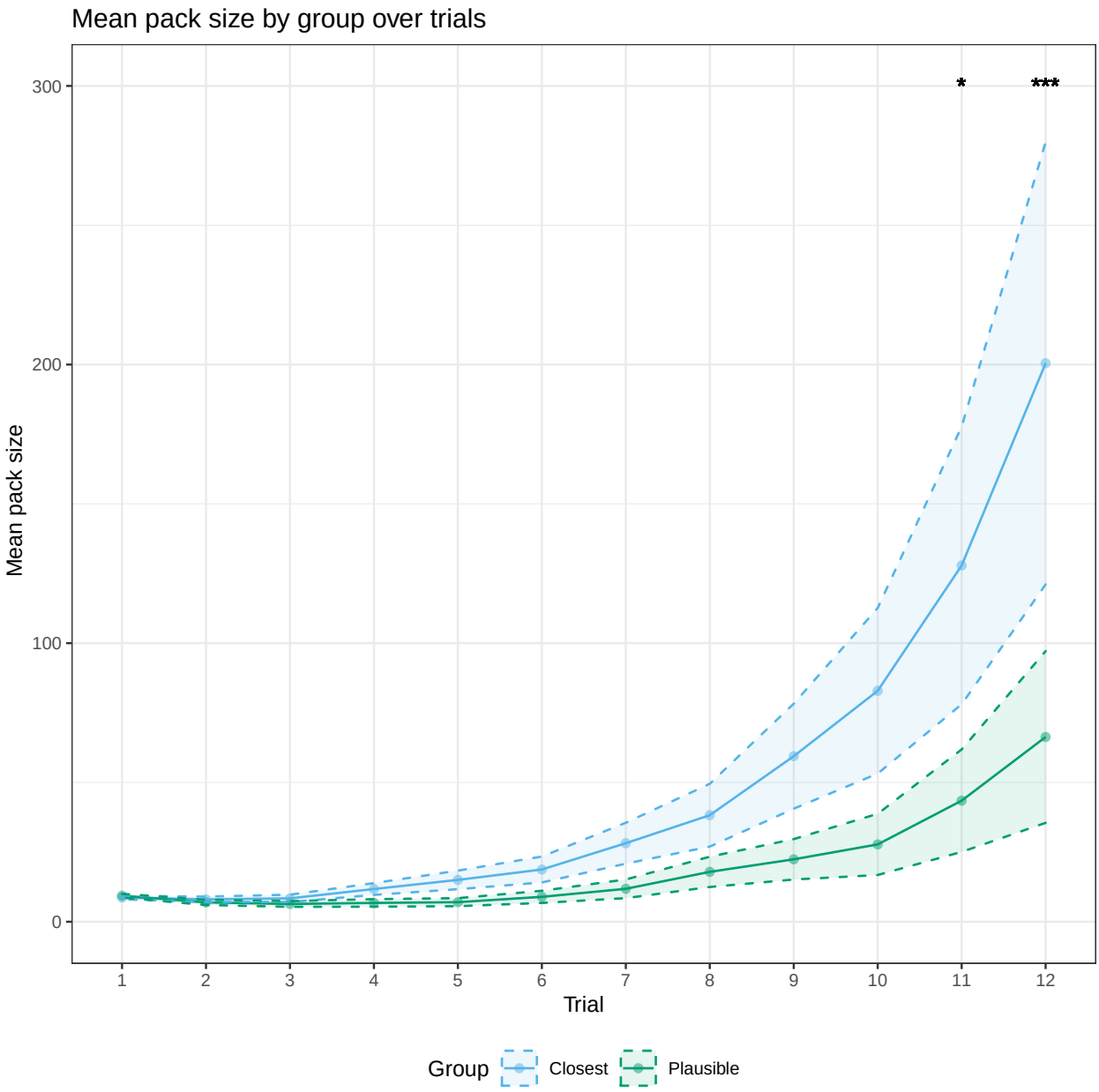
Results The analysis revealed:

- a significant interaction (group x trials): F(11,792.0000015)=2.1193007, p=0.0170773,η_{p}^2=0.0285931

Additionally:

- there was a significant main effect of trialnumber (time): F(11,792.0000015)=7.5850697, p=0,η_{p}^2=0.0953077
- however, there was a no main effect of group: F(1,71.9999969)=2.5857348, p=0.112207,η_{p}^2=0.034668

Posthoc analysis revealed significant differences betweeen groups in trials 11 (t(472)=4.020, p=0.0117) and 12 (t(472)=2.530, p=0.0001):



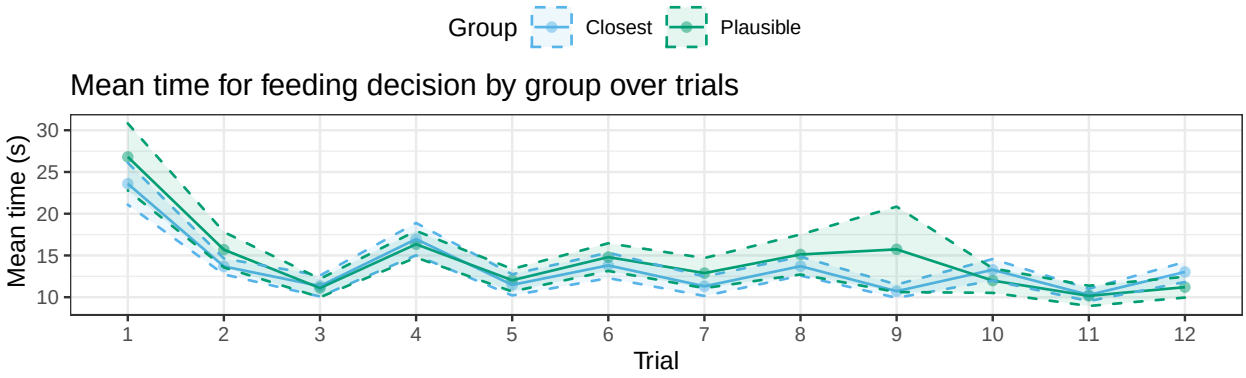
H1.2) Users in the plausible condition become quicker in deciding what plants to choose in the final blocks, because choice of the right plants will become more automatic

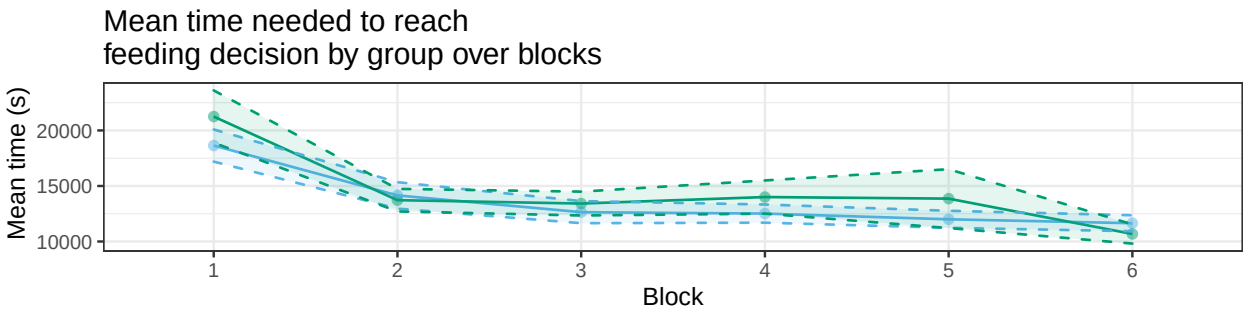
Again, first peek at the data: Descriptive stats + plotting the RT trajectories per trial and block for each person individually.

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

```
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1722    8426   11166   13599   15638   70031
##
## $P
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1388    7140   11022   14489   17876  180857
```

```
## [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             17990253   17990253     1    72  0.2347 0.62955
## TrialNr       12495620842 1135965531    11   792 14.8177 0.00000
## group:TrialNr   756052776   68732071    11   792  0.8966 0.54332
```

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

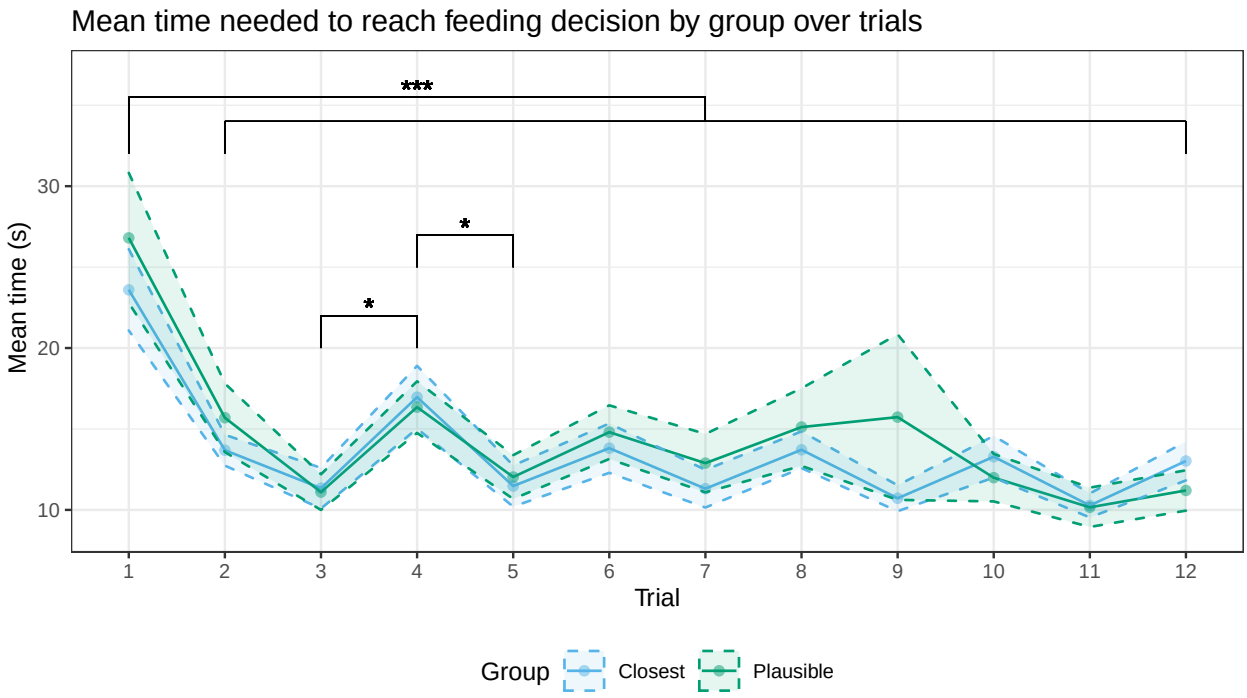
The analysis revealed:

- There was a highly significant main effect of trials (time): $F(11,792.0000013)=14.817721$, $p=0$,$\eta_p^2=0.1706762$

The other comparisons did not show significant differences:

- main effect of group: $F(1,71.9999967)=0.2346678$, $p=0.6295545$,$\eta_p^2=0.0032487$
- interaction (group x trials): $F(11,792.0000014)=0.8965524$, $p=0.5433182$,$\eta_p^2=0.012299$

Post-hoc analysis of the main effect of trial showed significant differences between trial 1 and all other trials (all $t(792)>5.90$, $p<0.0001$), between trial 3 and 4 ($t(792)=3.765$, $p=0.0118$) and between trials 4 and 5 ($t(792)=3.395$,$p=0.0476$).

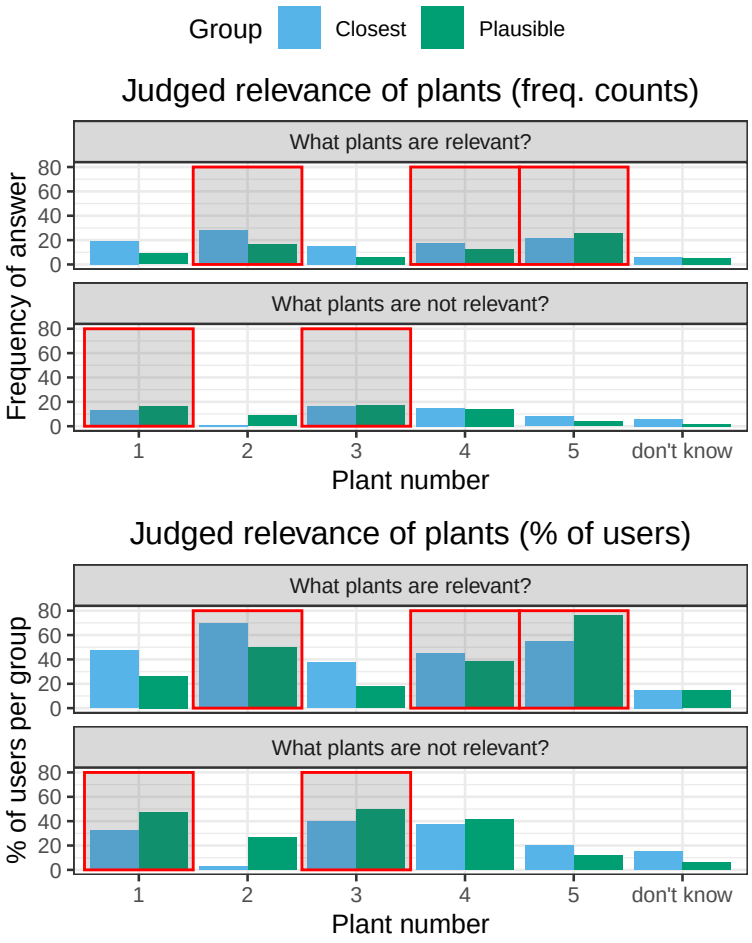


H1.3) Users in the plausible condition can more clearly state which plants were crucial for the Shubs to prosper (questionnaire items 1 and 2)

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

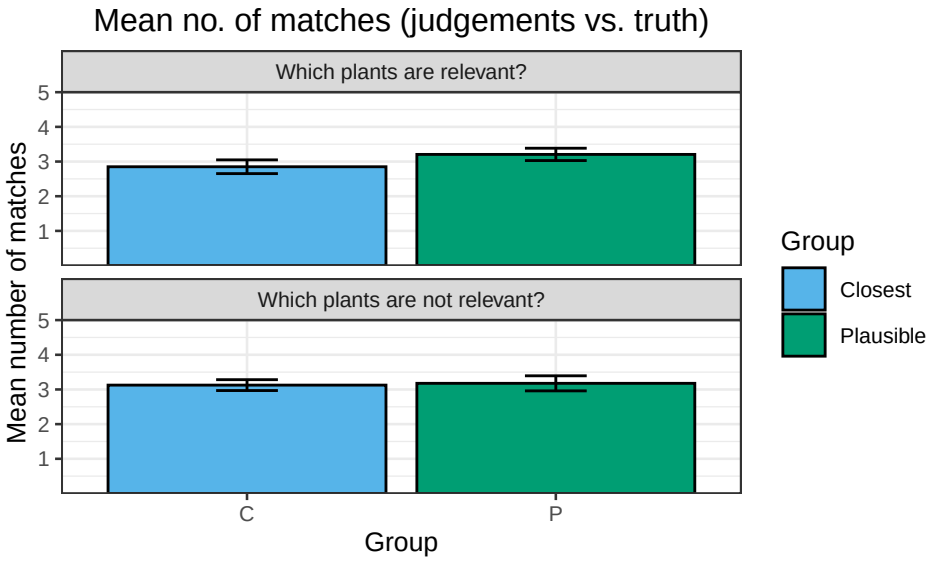
```
##     userId          group   itemNo  responseNo    checked
##  Length:888         C:480   1:444   1:148      Min.   :0.0000
##  Class :character   P:408   2:444   2:148      1st Qu.:0.0000
##  Mode  :character                   3:148      Median :0.0000
##                                     4:148      Mean   :0.3435
##                                     5:148      3rd Qu.:1.0000
##                                     6:148      Max.   :1.0000
```

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



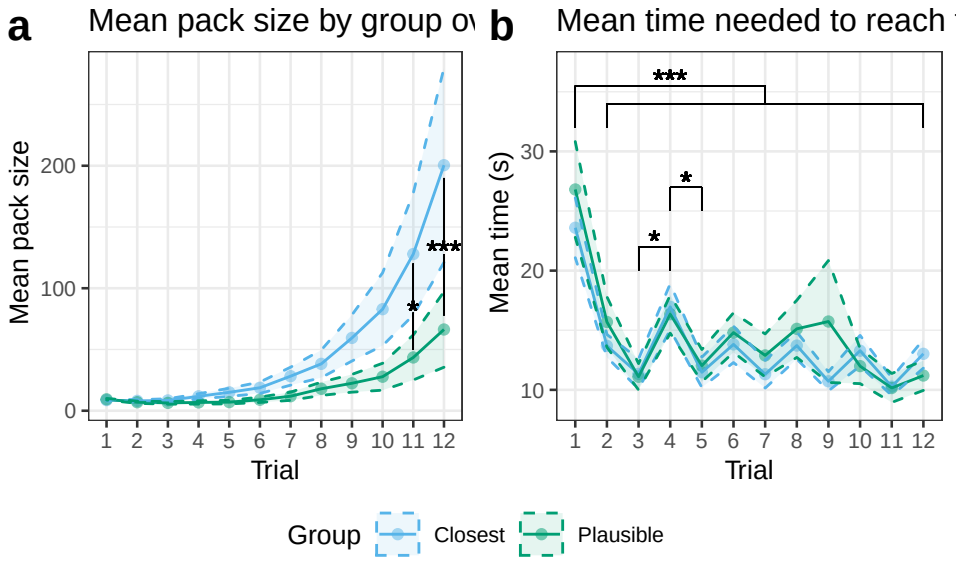
How to evaluate this statistically? Let's just count the matches between 'judged as relevant' / 'judged as irrelevant' user vectors and the true 'relevant' / 'irrelevant' factors.

```
## [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 = 3.2058824, SEM = 0.1780235) and users in the closest condition (M = 2.85, SEM = 0.1979057) using a Wilcox test. This showed
  - for wilcoxon test: U=781, p=0.2545969, r = 0.1324357
- 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 = 3.1764706, SEM = 0.2172203) and users in the closest condition (M = 3.125, SEM = 0.1568418) using a Wilcox test. This showed
  - for wilcoxon test: U=721.5, p=0.6426705, r = 0.0539352



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 make 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 9).

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."

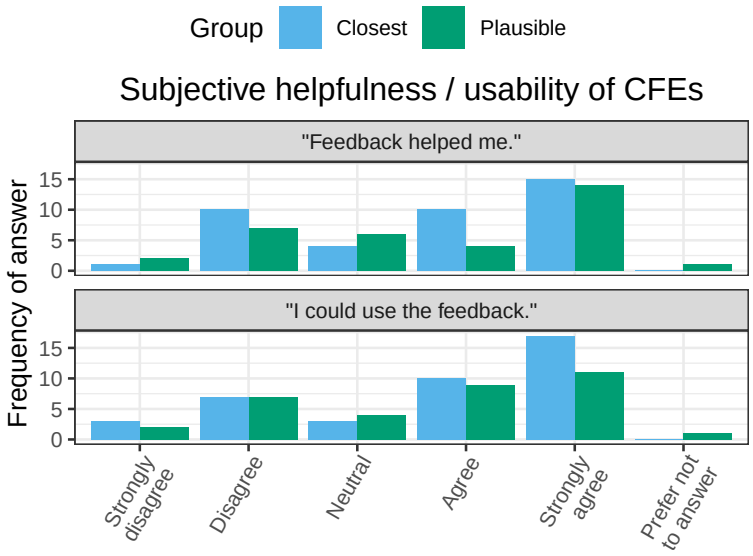
Item 6: "I was able to use the feedback based on what choice would have led to a better result to increase the number of Shubs."

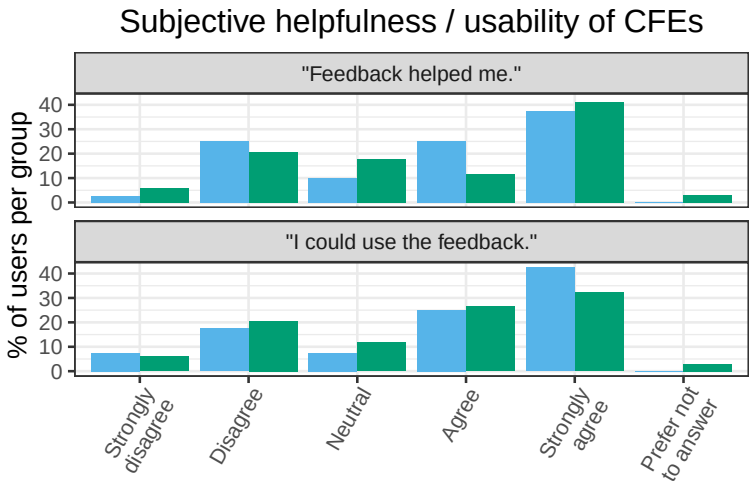
We will talk about these as quantifying how subjectively helpful (item 5) and how usable (item 6) they were.

```
##     userId          group   itemNo  responseNo    checked
##  Length:888         C:480   5:444   1:148      Min.   :0.0000
##  Class :character   P:408   6:444   2:148      1st Qu.:0.0000
##  Mode  :character                   3:148      Median :0.0000
```

```
##                                   4:148      Mean   :0.1667
##                                   5:148      3rd Qu.:0.0000
##                                   6:148      Max.   :1.0000
```

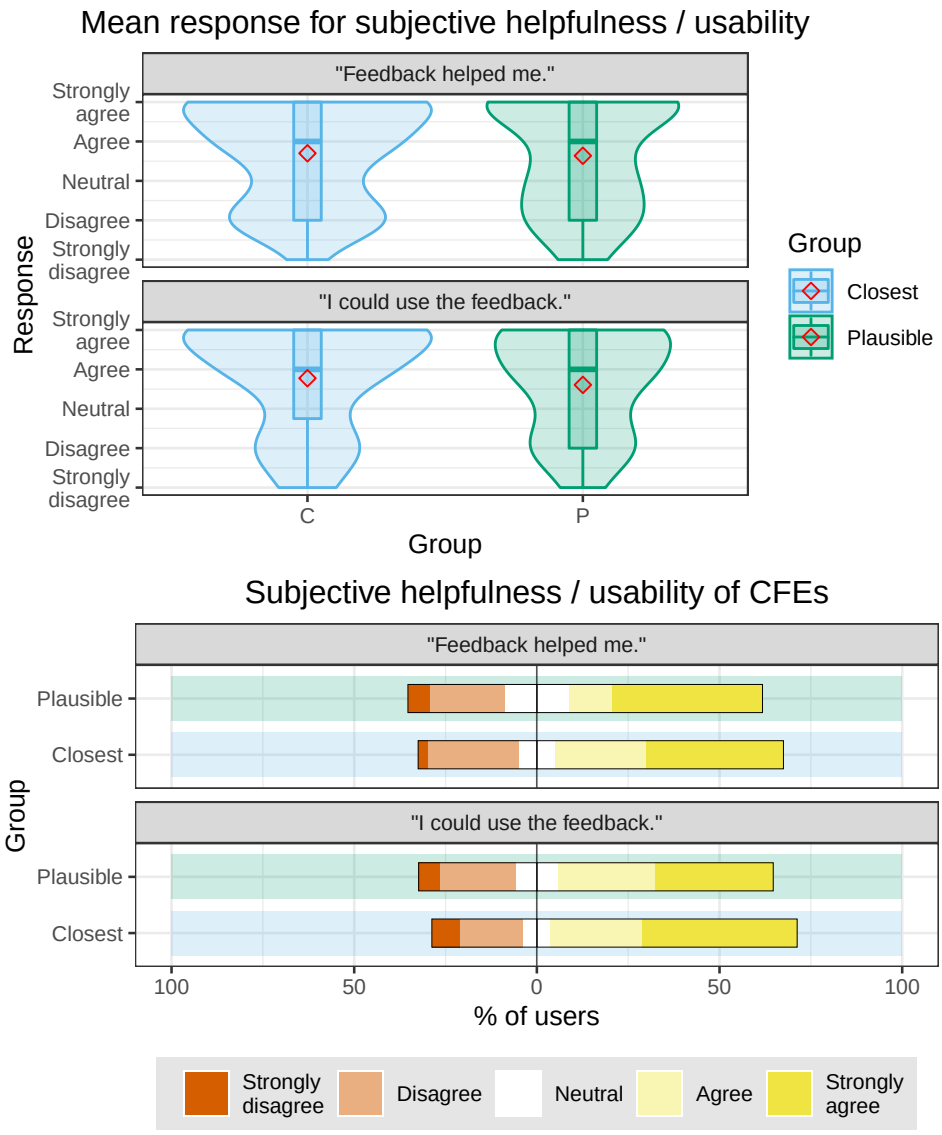
```
## [1] "Display figures showing user responses in relevant survey items:"
```





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 = 3.6363636, SEM = 0.2415942) and users in the closest condition (M = 3.7, SEM = 0.2031798) using a Wilcoxon–Mann–Whitney U test. This showed: U=656, p=0.9676533, r = -0.0047462
- Is there a significant difference in terms of subjective usability?: We compared responses for subjective usability for users in plausible condition (M = 3.6060606, SEM = 0.2300337) and users in the closest condition (M = 3.775, SEM = 0.2162842) using a Wilcoxon–Mann–Whitney U test. This showed: U=603, p=0.5133334, r = -0.0765047

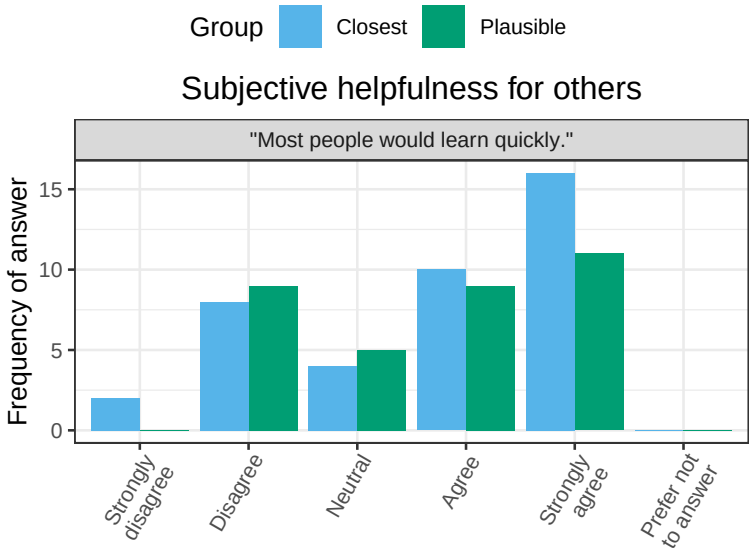
H2.2) Users imagine plausible CFEs to be more helpful for others users, too (questionnaire item 9).

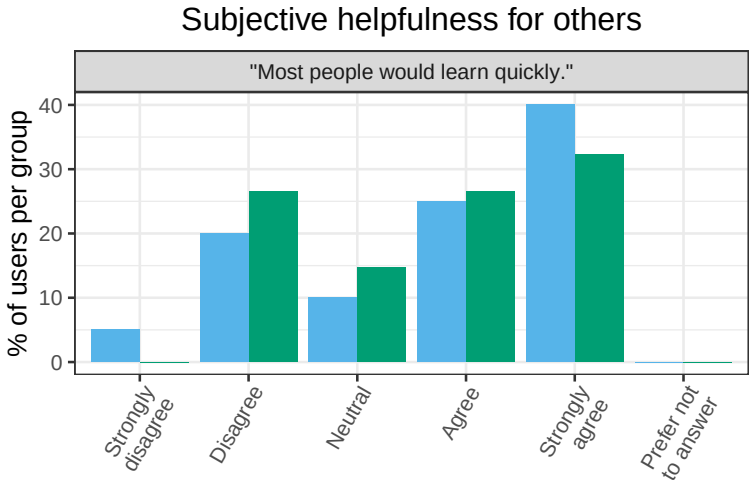
item 9: "I think most people would learn to work with the feedback on what choice would have led to a better result very quickly."

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:444         C:240   Min.   :9   1:74       Min.   :0.0000
##  Class :character   P:204   1st Qu.:9   2:74       1st Qu.:0.0000
##  Mode  :character           Median :9   3:74       Median :0.0000
##                             Mean   :9   4:74       Mean   :0.1667
##                             3rd Qu.:9   5:74       3rd Qu.:0.0000
##                             Max.   :9   6:74       Max.   :1.0000
```

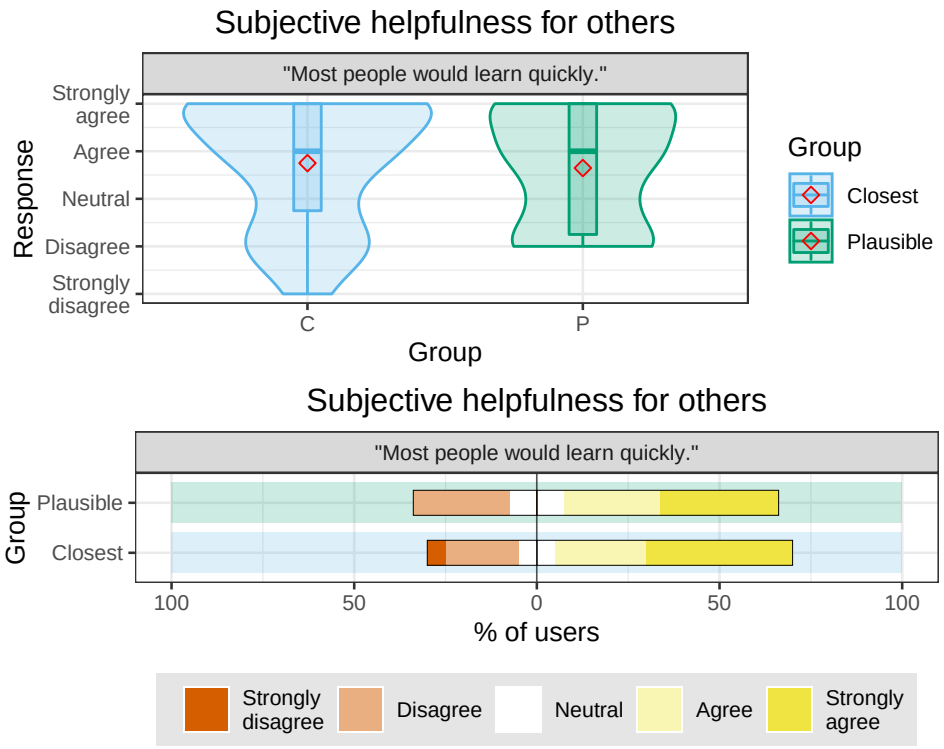
```
## [1] "Display figures showing user responses in relevant survey items:"
```





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 = 3.6470588, SEM = 0.2063275) and users in the closest condition (M = 3.75, SEM = 0.2080126) using a Wilcoxon–Mann–Whitney U test. This showed: U=637, p=0.6309341, r = -0.0558468

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:

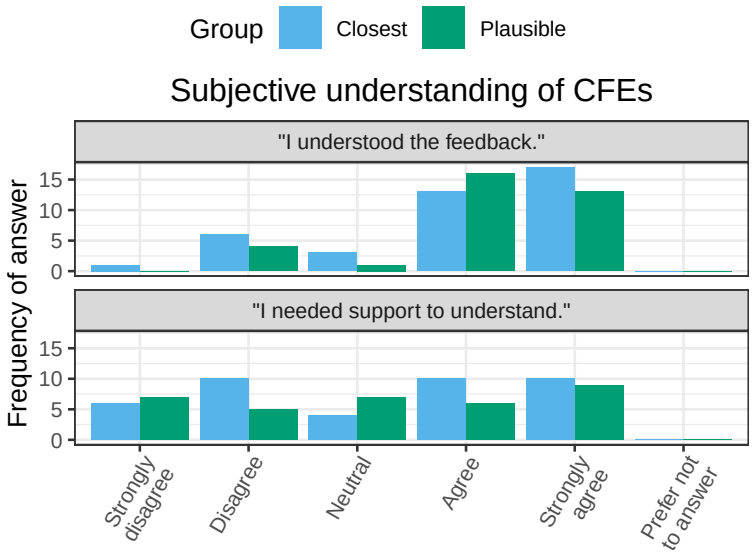
H3) We do not expect users in different conditions to differ in terms of how well they understood the explanations per se, or needing support for understanding, because explanations are basically the same structurally (questionnaire items 3, 4).

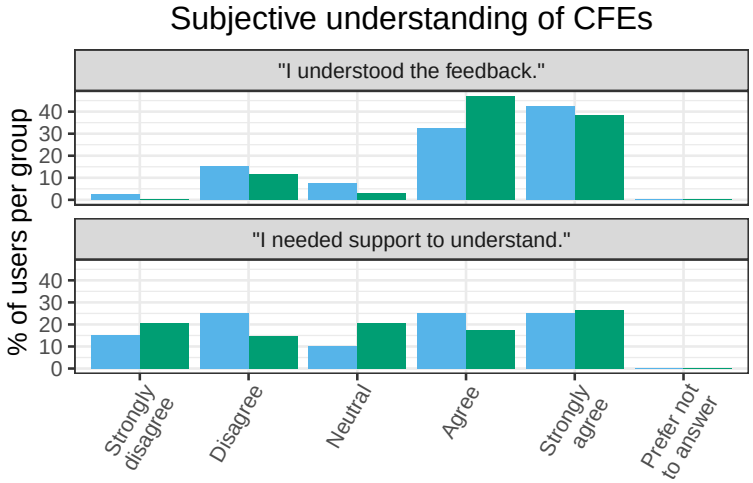
Item 3: "I understood the feedback on what choice would have led to a better result."

Item 4: "I needed support to understand the feedback on what choice would have led to a better result."

```
##     userId          group   itemNo  responseNo    checked
##  Length:888         C:480   3:444   1:148      Min.   :0.0000
##  Class :character   P:408   4:444   2:148      1st Qu.:0.0000
##  Mode  :character                   3:148      Median :0.0000
##                                     4:148      Mean   :0.1667
##                                     5:148      3rd Qu.:0.0000
##                                     6:148      Max.   :1.0000
```

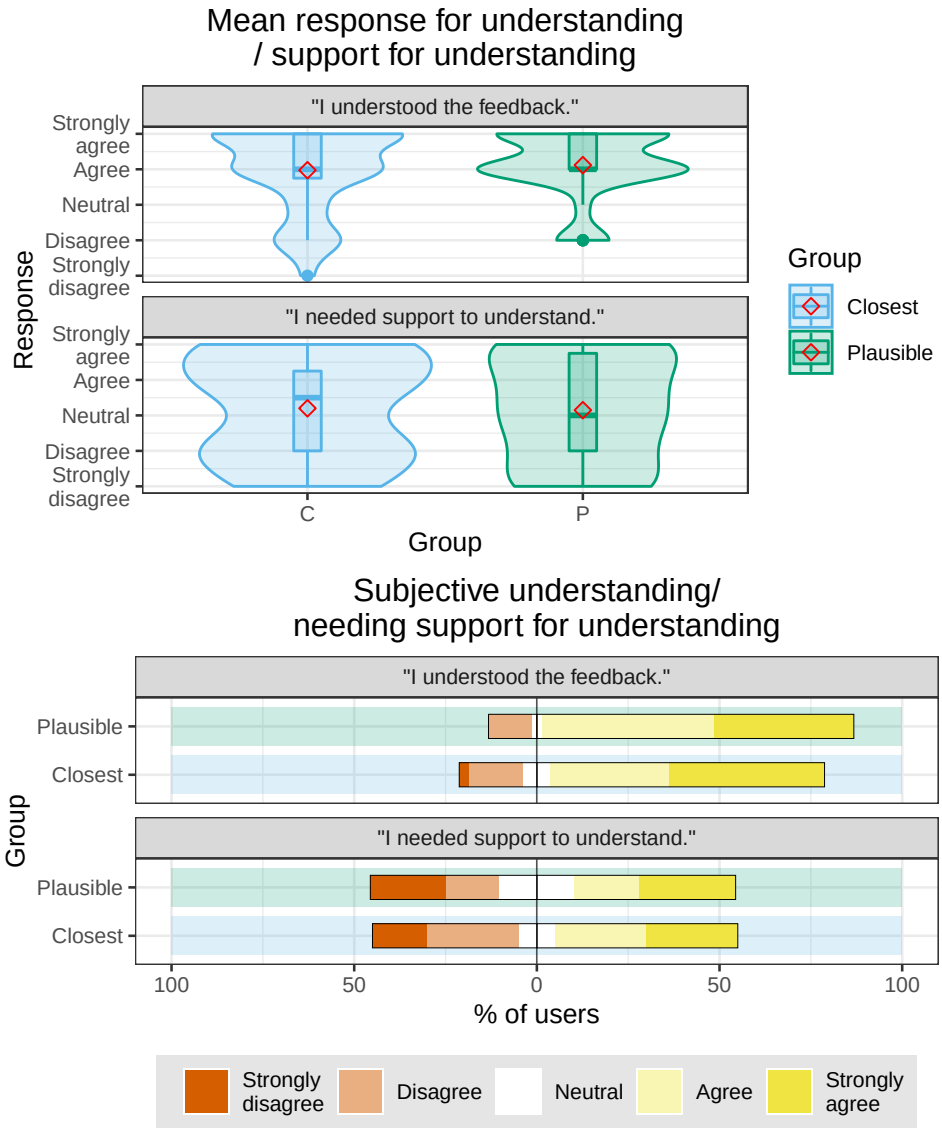
```
## [1] "Display figures showing user responses in relevant survey items:"
```





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 = 4.1176471, SEM = 0.1622299) and users in the closest condition (M = 3.975, SEM = 0.1842779) using a Wilcoxon–Mann–Whitney U test. This showed: U=773.5, p=0.2002938, r = 0.1488801
- Is there a significant difference in terms of needing support to understand explanations?: We compared responses of users in plausible condition (M = 3.1470588, SEM = 0.2572734) and users in the closest condition (M = 3.2, SEM = 0.2298271) using a Wilcoxon–Mann–Whitney U test. This showed: U=667, p=0.8897197, r = -0.0161188

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

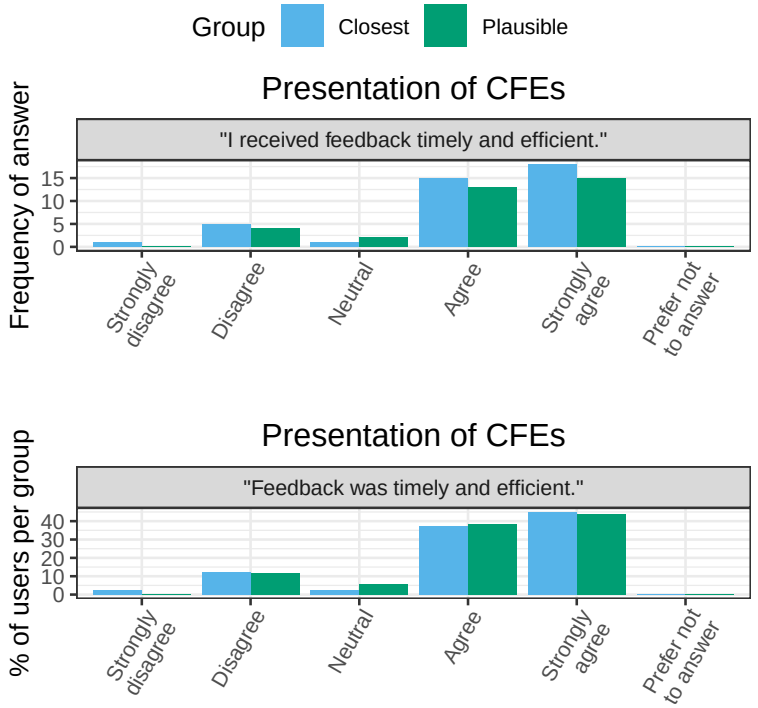
H4) We expect timing and efficacy of how CFEs were presented to be comparable, as it was literally the same (questionnaire item 10) - a further control.

Item 10: “I received the feedback on what choice would have led to a better result in a timely and efficient

manner."

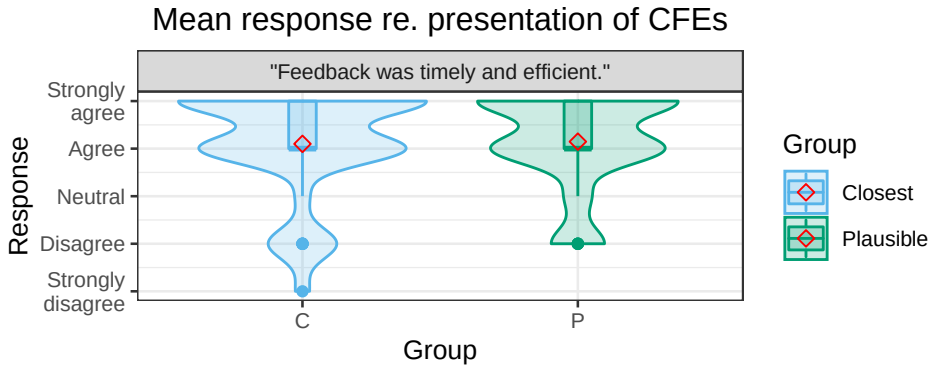
```
##     userId           group       itemNo     responseNo    checked
##  Length:444         C:240   Min.   :10   1:74       Min.   :0.0000
##  Class :character   P:204   1st Qu.:10   2:74       1st Qu.:0.0000
##  Mode  :character           Median :10   3:74       Median :0.0000
##                             Mean   :10   4:74       Mean   :0.1667
##                             3rd Qu.:10   5:74       3rd Qu.:0.0000
##                             Max.   :10   6:74       Max.   :1.0000
```

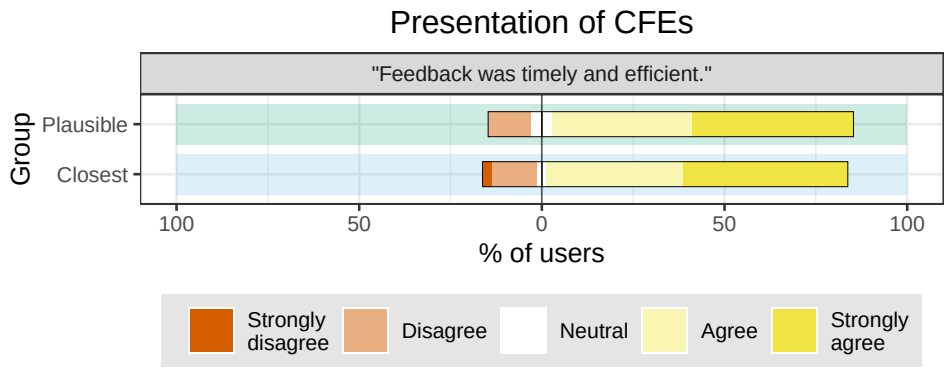
```
## [1] "Display figures showing user responses in relevant survey items:"
```



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 = 4.1470588, SEM = 0.1695772) and users in the closest condition (M = 4.1, SEM = 0.1746792) using a Wilcoxon–Mann–Whitney U test. This showed: U=680.5, p=1, r = 0

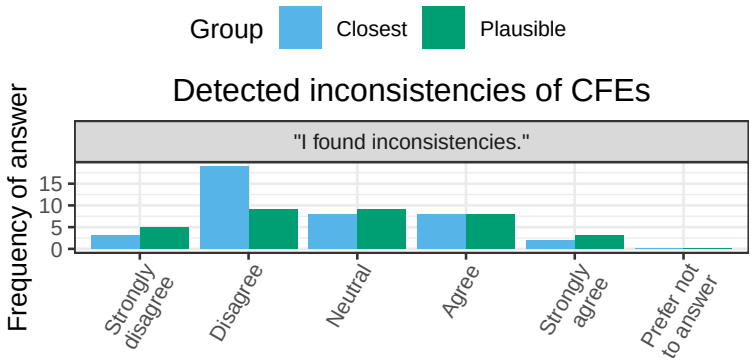
Final exploratory analysis

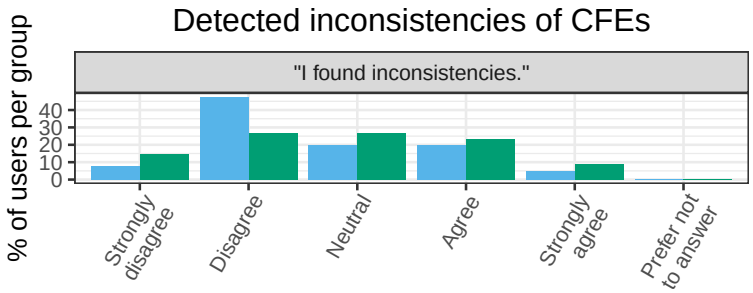
It is not clear whether users were uncover inconsistencies in the feedback. Maybe that is the case for "closest" CFEs when we're in the areas of "no training data"? Let's see what users responded.

Item 8: "I found inconsistencies in the feedback on what choice would have led to a better result."

```
##     userId          group       itemNo  responseNo    checked
##  Length:444         C:240   Min.   :8   1:74       Min.   :0.0000
##  Class :character   P:204   1st Qu.:8   2:74       1st Qu.:0.0000
##  Mode  :character           Median :8   3:74       Median :0.0000
##                             Mean   :8   4:74       Mean   :0.1667
##                             3rd Qu.:8   5:74       3rd Qu.:0.0000
##                             Max.   :8   6:74       Max.   :1.0000
```

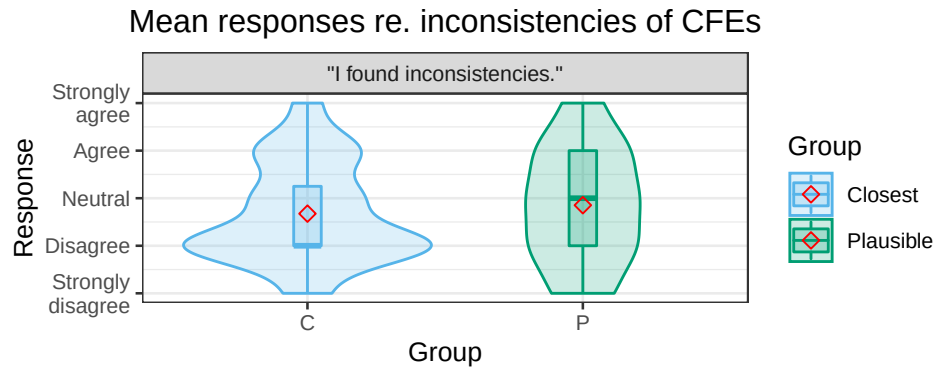
```
## [1] "Display figures showing user responses in relevant survey items:"
```

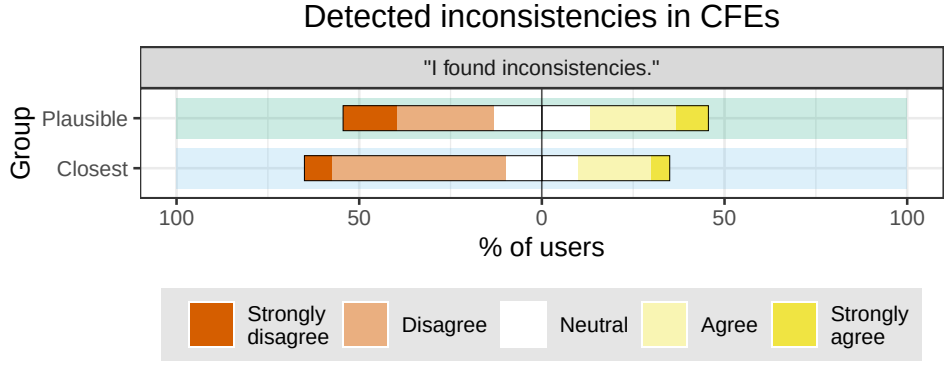




Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

```
## [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 = 2.8529412, SEM = 0.2074046) and users in

the closest condition (M = 2.675, SEM = 0.1655895) using a Wilcoxon–Mann–Whitney U test. This showed: U=743, p=0.4802334, r = 0.0820624

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.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.