Introducing the Alien Zoo: Summary of evaluation and results for Exp1 (DS4 condition, March 2022)

Contents

Introduction

This is an analysis of data acquired in the "Introducing the Alien Zoo" study run on Amazon mechanical turk in March 2022. 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 either counterfactual explanations (CFEs) regarding past choices, or no explanation. Computed CFEs correspond to "Control" counterfactuals that fulfill the "smallest feature change" condition (Wachter, Mittelstadt, and Russell 2017). This script evaluates data from Experiment 1 (three features, plant 2, 4, and 5, impacting growth rate).

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/IAZ/alienzoo/
```

General infos after removal of incomplete datasets

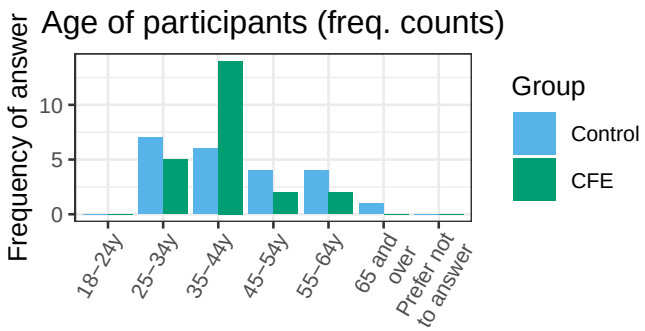
How many users do we have in our performance df before any cleaning (i.e., also including users with incomplete datasets)? 58

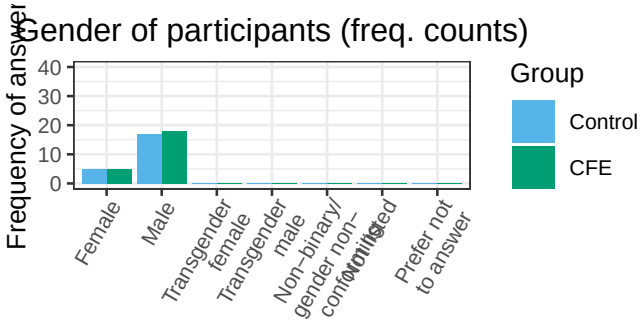
After cleaning, we have 45 participants. Of those,

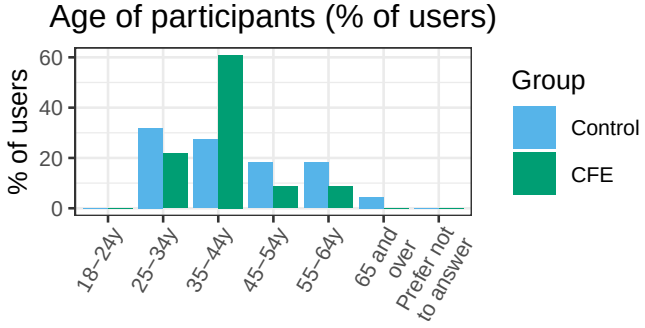
- 22 participants were in the control condition and
- 23 participants in the explanation condition.

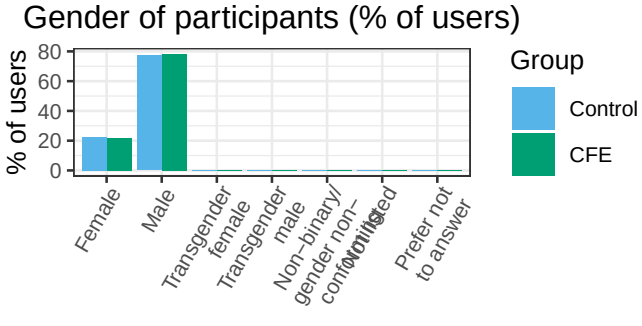
Check covariates across groups

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 45 participants, with 22 users in the control group (5 female, 17 male, median age group is 35-44years), and 23 users in the explanation group (5 female, 18 male, median age group is 35-44years).

The analysis showed for *Age*:

- We have age information for 23 users in the explanation and 22 users in the control group.
- Is there a significant difference in terms of age between the groups? We compared ages of users in explanation condition and users in the control condition using a Wilcox test. This showed: U=225.5, p=0.5155449, r = -0.0969303

The analysis showed for *Gender*:

- We have gender information for 23 users in the explanation and 22 users in the control group.
- Is there a significant difference in terms of gender between the groups? We compared gender distribution for users in explanation condition and users in the control condition using a Wilcox test. This showed
  - for wilcoxon test: U=255.5, p=0.9497276, r = 0.0093988

Quality criteria

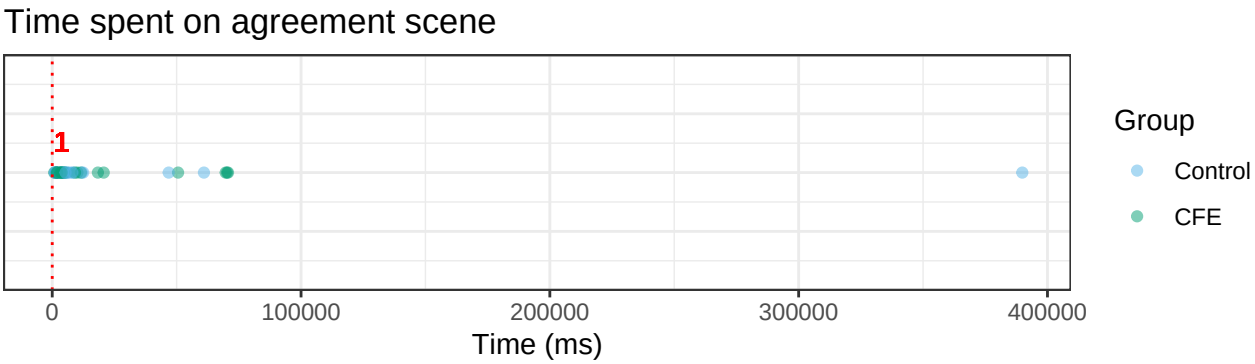
Before going into the hypotheses, we should apply some quality criteria to our data. Sub-quality data should be removed. The following subsections take care of such cases.

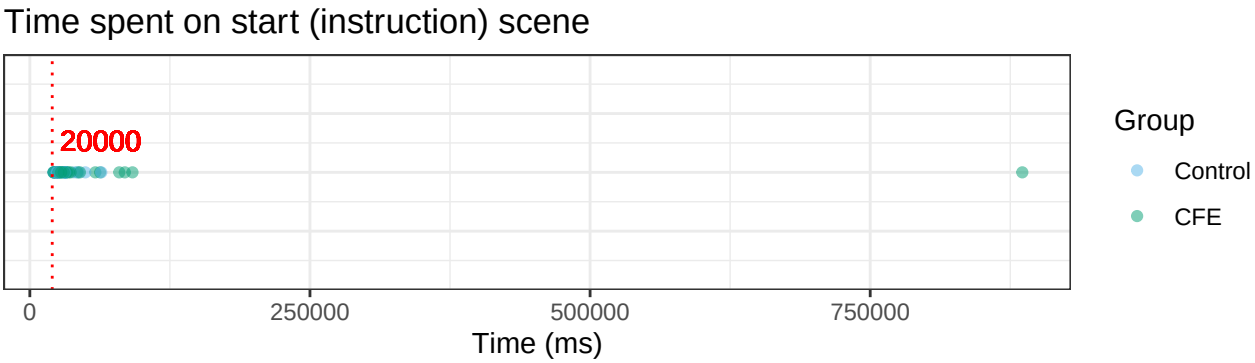
Identify "speeders"

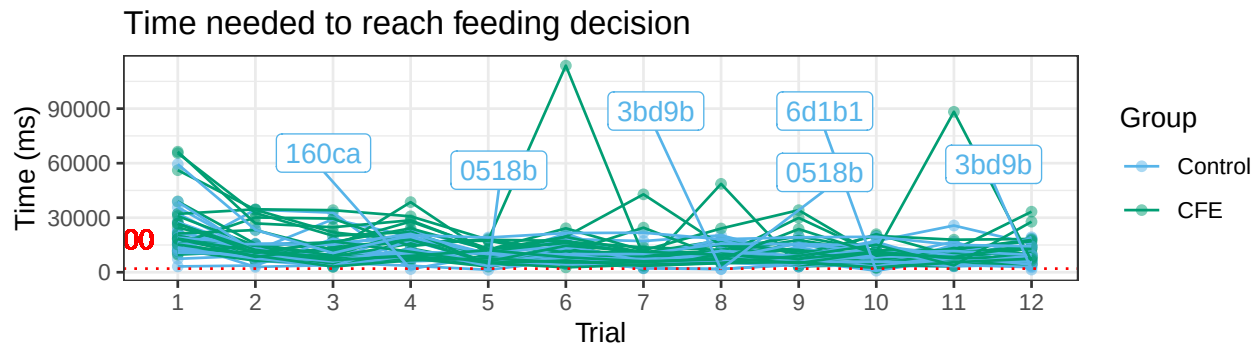
Speeders are people clicking through the study way too quickly to do the task properly.

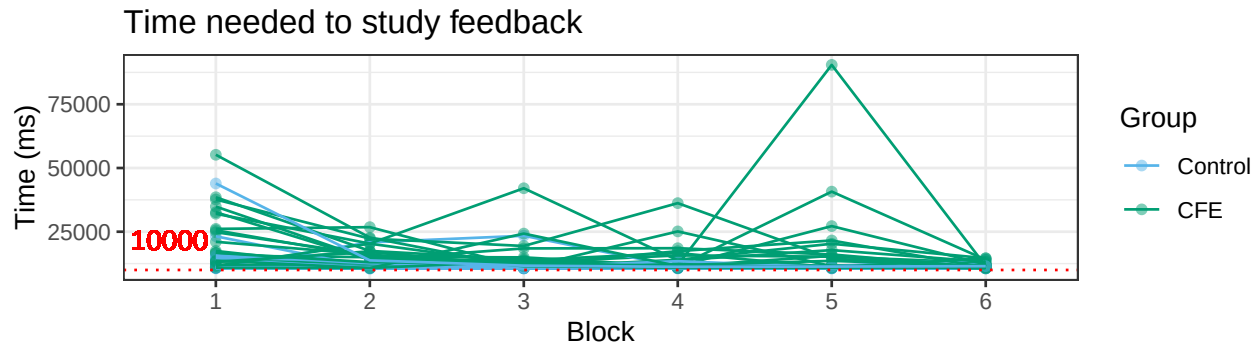
Aim: identify IDs being faster than specified values (variable per game part). This part will tag users that needed less than 2000ms to reach a feeding decision (suspiciously quick) in 4 or more trials.

```
## [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 58 users to begin with
- we remove 13 users that have incomplete datasets (aborted prematurely)
- we remove 0 whose information was not logged properly ** Note here: for one user, logging failed for the survey items 1 and 2 - we will keep data for this person anyway for all other measurements
- we remove 0 speeders
- we remove 2 users that failed both attention tests during the game
- we remove 0 users that failed the attention test in the survey
- we remove 4 users that straightlined in the game, despite not improving
- remove 0 users that straightlined in the survey

Finally: How many users do we have in our clean performance df? 39

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

Final, clean dataset

To sum up, in our final data we have 39 users, with 19 users in the control group (4 female, 15 male, median age group is 35-44years), and 20 users in the explanation group (5 female, 15 male, 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 20 users in the explanation and 19 users in the control group.
- Is there a significant difference in terms of age between the groups? We compared age of users in the explanation condition and users in the control condition using a Wilcox test. This showed: U=143, p=0.1680159, r = -0.2207538

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

- We have gender information for 20 users in the explanation and 19 users in the control group.
- Is there a significant difference in terms of gender between the groups? We compared gender distribution of users in explanation condition and users in the control condition using a Wilcox test. This showed
  - for wilcoxon test: U=182.5, p=0.7875987, r = -0.0431433

Hypotheses

The main hypothesis is the following:

H1) CFEs will help users tasked to discover unknown relationships in data. We expect this to affect objective as well as subjective understandability.

That means, we expect users in the explanation 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 (survey 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 feedback (CF-style explanation vs. overview over past choices only) useful, and in how far they could made use of it, with an advantage of providing CFEs (survey items 5, 6)
- H2.2) Users imagine that providing CFEs to be more helpful for others users, too (survey item 9).

Moreover:

H3) We expect users in different conditions not to differ in terms of how well they understood the feedback per se, or needing support for understanding. This would be good, because it means that the added information provided does not overload the participant's cognitive capacities. (survey 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 feedback was provided to be comparable (survey item 10) - a further control. Could still be that there is a difference here, as less useful feedback (control) is perceived less efficient.

H5) Finally, we predict that users will not have uncoverred inconsistencies in the feedback. It would be weird for the control group; and the models were really good, and we trust CEML to do a good job. (survey item 8)

Statistical assessment

[...] Comparisons of performance over time between users in the explanation and control 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 the 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).

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: Providing CFEs helps users

Recap the full hypothesis:

H1) CFEs will help users tasked to discover unknown relationships in data. We expect this to affect objective as well as subjective understandability.

That means, we expect users in the explanation 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 (survey items 1 and 2).

H1.1) Users in the explanation 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 (figure not shown in .pdf).

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

```
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     2.0     2.0     3.5     2.0    14.0
##
## $E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.00    2.00    9.40   11.25   93.00
```

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       4126.3 375.11    11   407 15.7581 0.000000
## group           89.4  89.38     1    37  3.7547 0.060318
## trialNo:group 1741.0 158.28    11   407  6.6490 0.000000
```

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

Results The analysis revealed:

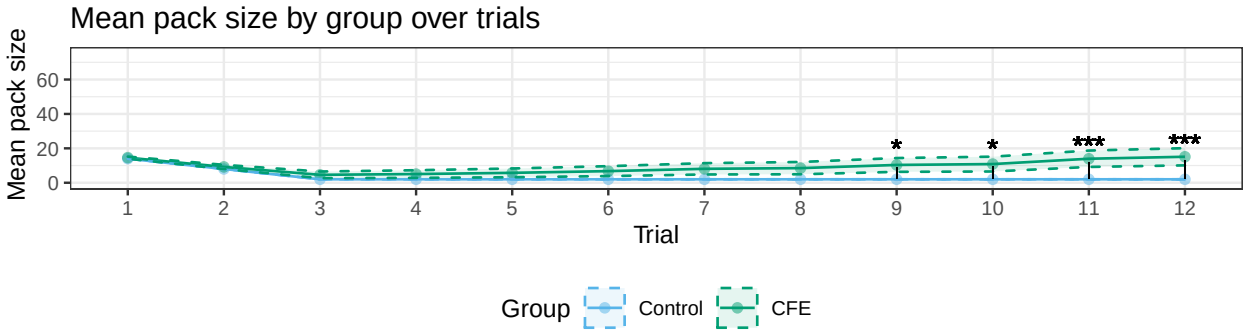
- a significant interaction (group x trials): F(11,407.0000001)=6.6489606, p=0,η_p^2=0.152328

Additionally:

- there was a significant main effect of trialnumber (time): F(11,407.0000001)=15.7580944, p=0,η_p^2=0.2986858
- however, there was a no main effect of group: F(1,36.9999999)=3.754748, p=0.0603185,η_p^2=0.0921303 (mean ShubNo explanation group: 9.4, sem=0.9590427; mean ShubNo control group: 3.5, sem=0.2369966).

Posthoc analysis revealed significant differences betweeen groups from trial 9 onwards (trial 9: t(56.7)=2.461, p=0.0169, Cohen's d=-1.711; trial 10: t(56.7)=2.609, p=0.0116, Cohen's d=-1.814; trial 11: t(56.7)=3.522, p=0.0009, Cohen's d=-2.449; trial 12: t(56.7)=3.861, p=0.0003, Cohen's d=-2.685):

```
## [1] "Display figure showing development of pack size over trials / blocks:"
```



H1.2) Users in the explanation 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 (plots generated, but not shown in final .pdf).

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

```
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1338    5748    9108   10537   12948   59454
##
## $E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2262    6629   10484   13557   16614  113671
```

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           245041200 245041200     1    37  3.9761 0.05356
## TrialNr        8829677156 802697923    11   407 13.0249 0.00000
## group:TrialNr   654449466  59495406    11   407  0.9654 0.47748
```

```
## 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(11,407.0000007)=13.0249132, p=0,η_p^2=0.2603685

The other main effect of group did not reach significance * ... and also a significant main effect of group: F(1,36.9999985)=3.9761413, p=0.0535583,η_p^2=0.0970355 (mean RT explanation group: 13.55735s, sem=0.7618599s; mean RT control group: 10.5368991s, sem=0.4667599s).

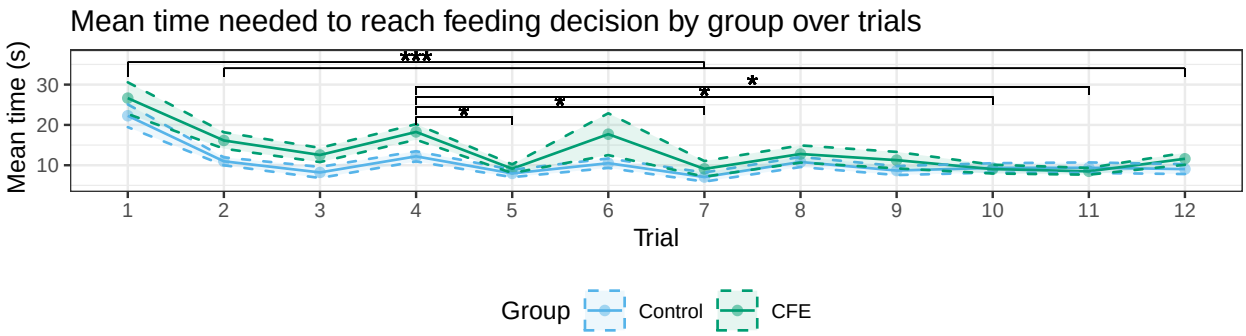
- ... and neither did the interaction::
- interaction (group x trials): F(11,407.0000006)=0.9653974, p=0.4774818,η_p^2=0.0254284

Post-hoc analysis of the main effect of trial showed significant differences between

- trial 1 and all other trials (all t(407)>=5.189, p<.0001, Cohen's d=1.17536 or higher).
- between trial 4 and trial 5 (t(407)=3.755, p=0.0131, Cohen's d=0.85056)
- between trial 4 and trial 7 (t(407)=4.020, p=0.0046, Cohen's d=0.91069)
- between trial 4 and trial 10 (t(407)=3.397, p=0.0494, Cohen's d=0.76949)

- between trial 4 and trial 11 (t(407)=3.537, p=0.0298, Cohen's d=0.80129)

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



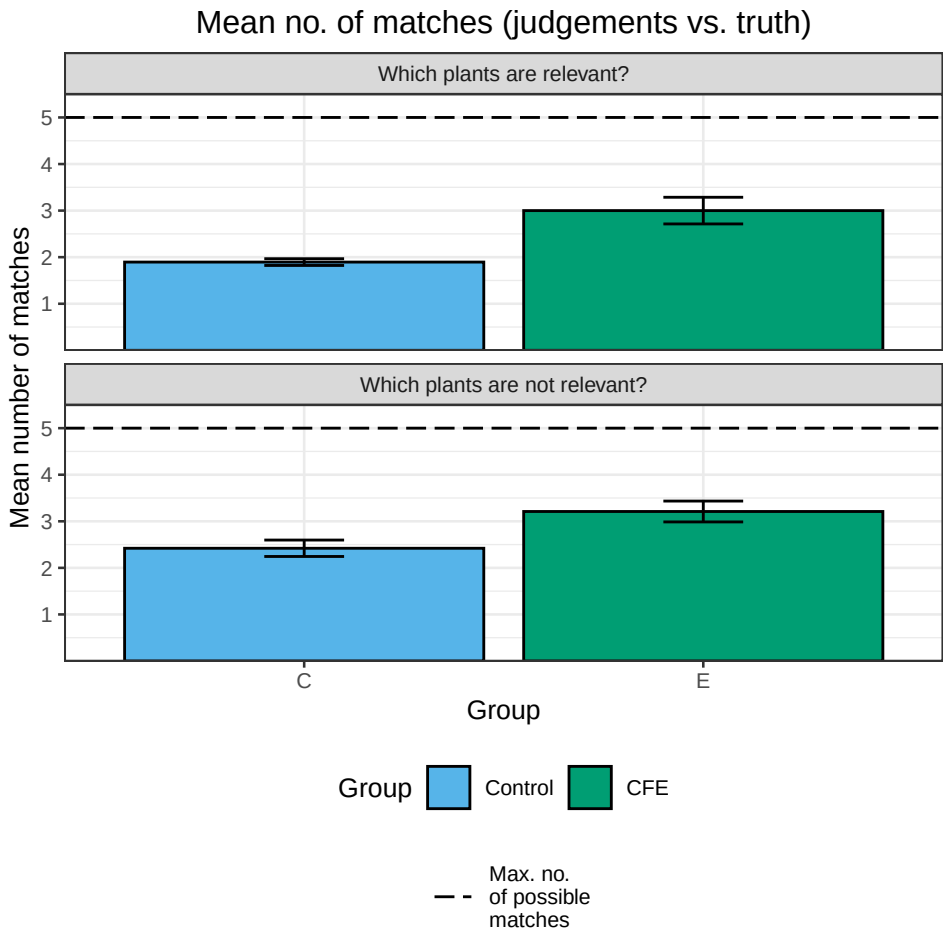
H1.3) Users in the explanation condition can more clearly state which plants were crucial for the Shubs to prosper (survey items 1 and 2)

survey 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:456         C:228   1:228   1:76       Min.   :0.0000  
##  Class :character   E:228   2:228   2:76       1st Qu.:0.0000  
##  Mode  :character                   3:76       Median :0.0000  
##                                     4:76       Mean   :0.3048  
##                                     5:76       3rd Qu.:1.0000  
##                                     6:76       Max.   :1.0000
```

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.

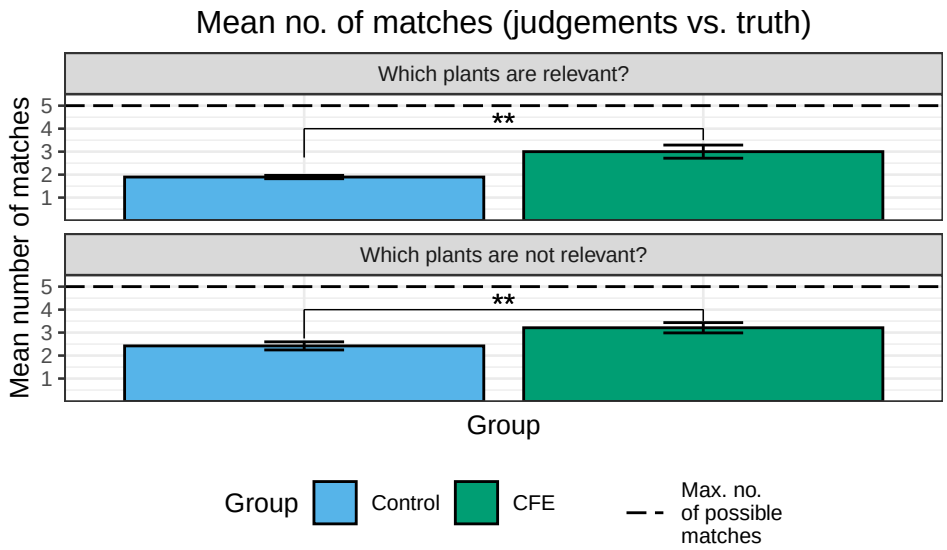
```
## [1] "Mean number of matches between user judgements and ground truth for relevant and irrelevant pla
```



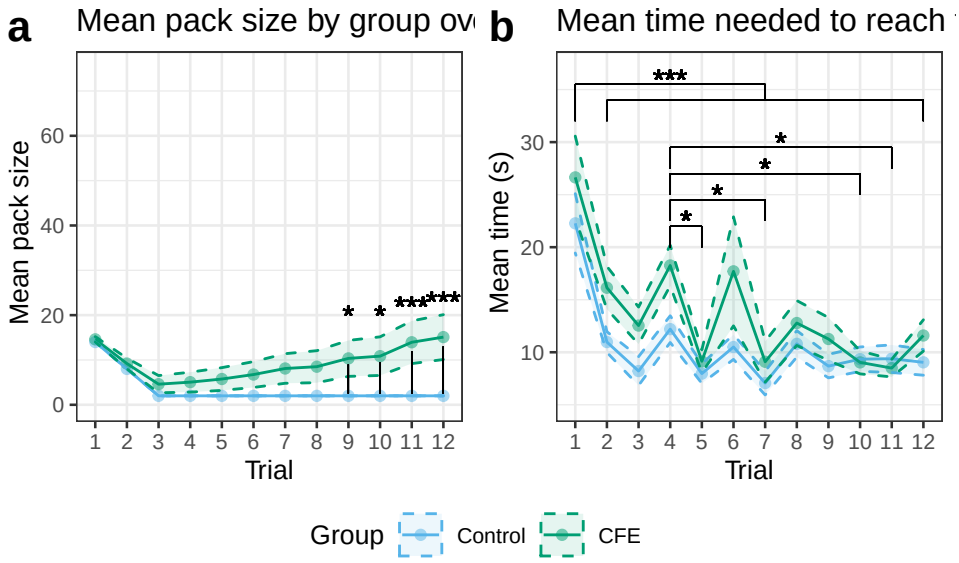
The analysis revealed:

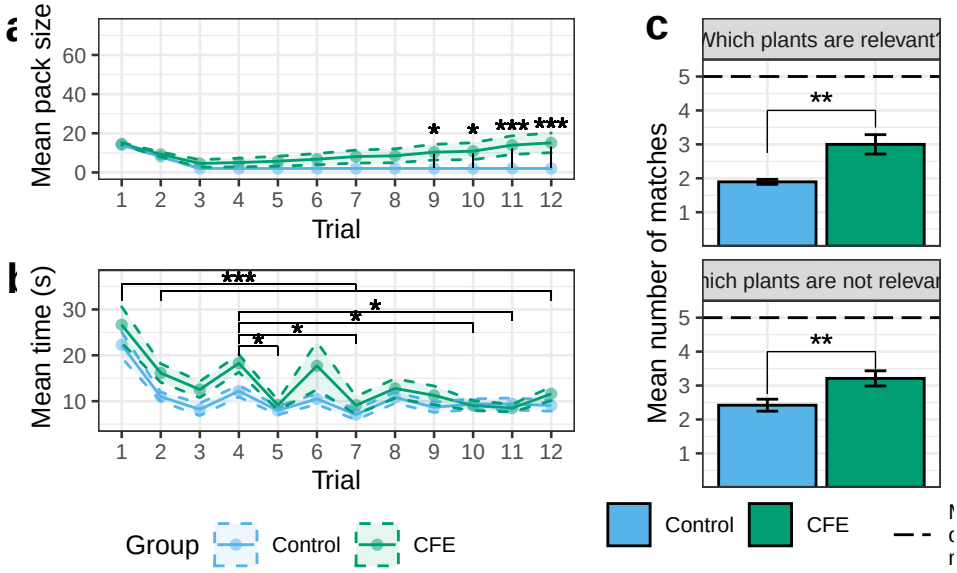
- Is there a significant difference in terms of matches between plants judged as relevant and ground truth?: We compared number of matches for users in explanation condition (M = 3, SEM = 0.2861317) and users in the control condition (M = 1.8947368, SEM = 0.0723352) using a Wilcox test. This showed
  - for wilcoxon test: U=281.5, p=0.0014243, r = 0.5174332
  - So YES! People receiving explanations could better identify relevant features.
  - Note: a large proportion on the control group replied with "I don't know," also indicating the troubles these participants had.
- 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 explanation condition (M = 3.2105263, SEM = 0.223985) and users in the control condition (M = 2.4210526, SEM = 0.1763136) using a Wilcox test. This showed
  - for wilcoxon test: U=264.5, p=0.0092607, r = 0.4221449

```
## [1] "Display figures showing matches between user responses and ground truth in relevant survey items
```



H1) Final plot for publication





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 providing CFEs (survey items 5, 6)
- H2.2) Users imagine CFEs to be more helpful for others users, too (survey 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 providing CFEs (survey items 5, 6)

Diving more into survey results.

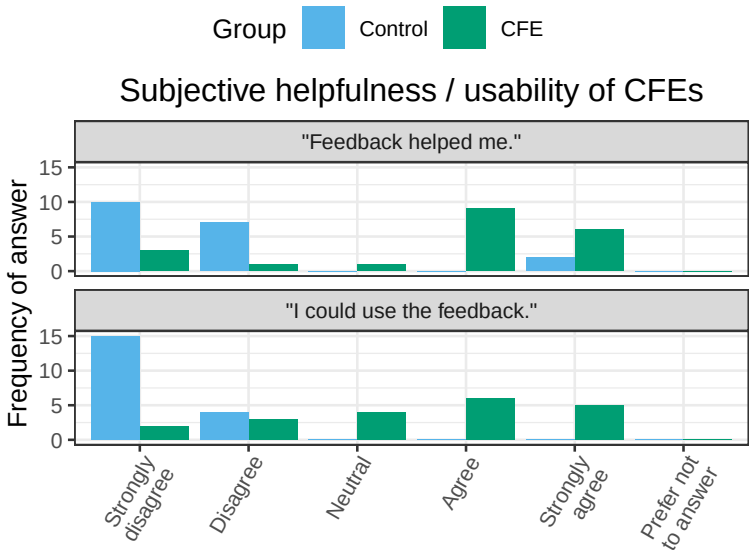
Item 5: "I found that the feedback on what choice would have led to a better result helped me to increase the number of Shubs."

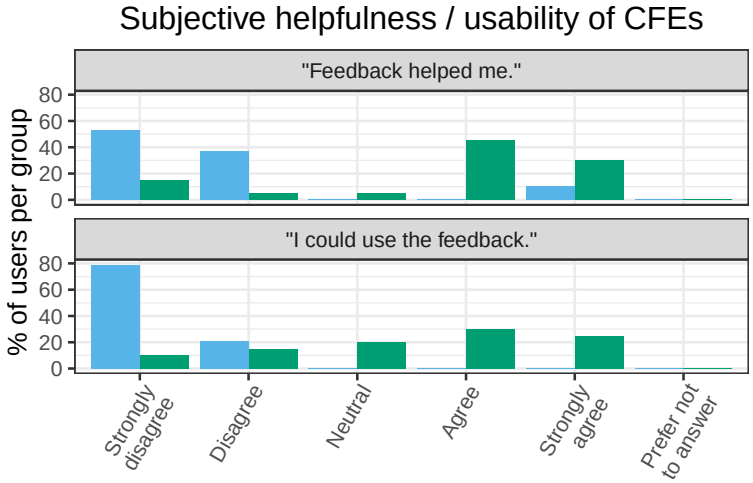
Item 6: "I was able to use the feedback based on what choice would have led to a better result to increase the number of Shubs."

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:468         C:228   5:234   1:78       Min.   :0.0000
##  Class :character   E:240   6:234   2:78       1st Qu.:0.0000
##  Mode  :character                   3:78       Median :0.0000
##                                     4:78       Mean   :0.1667
##                                     5:78       3rd Qu.:0.0000
##                                     6:78       Max.   :1.0000
```

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



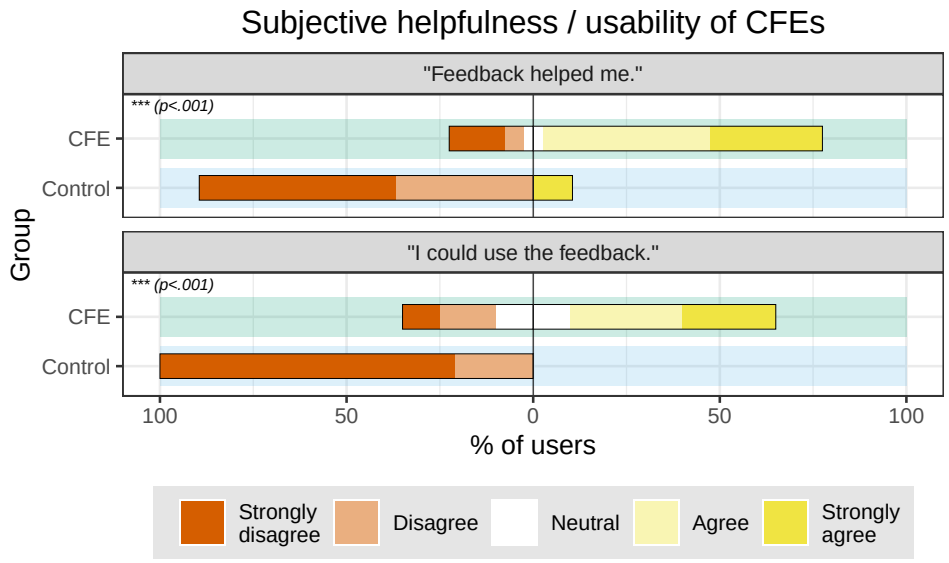


On to the statistical comparison: for Likert-scale, we want a non-parametric statistical test for ordinal data, that's the Wilcoxon–Mann–Whitney U test.

The analysis revealed:

- Is there a significant difference in terms of subjective helpfulness between groups? We compared responses for subjective helpfulness for users in explanation condition (M = 3.7, SEM = 0.3086473) and users in the control condition (M = 1.7894737, SEM = 0.2817961) using a Wilcoxon–Mann–Whitney U test. This showed: U=306.5, p=0.0007449, r = 0.5400316
- Is there a significant difference in terms of subjective usability?: We compared responses for subjective usability for users in explanation condition (M = 3.45, SEM = 0.2944665) and users in the control condition (M = 1.2105263, SEM = 0.0960917) using a Wilcoxon–Mann–Whitney U test. This showed: U=351, p=0.0000021, r = 0.7590116

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



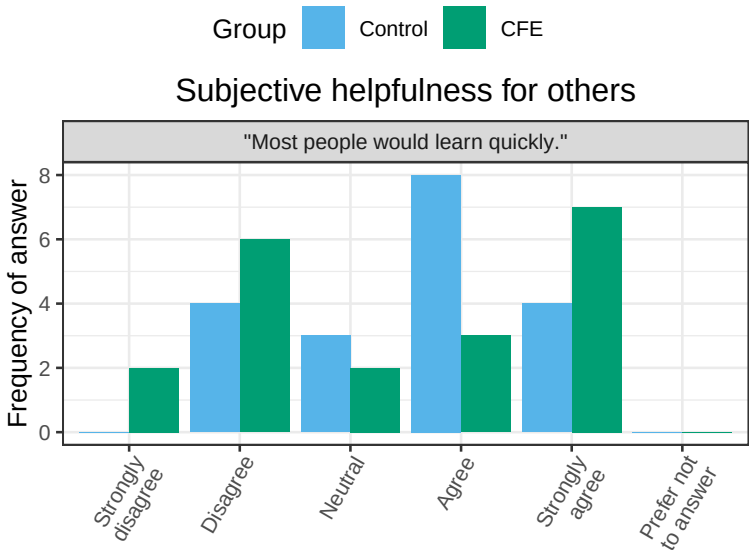
H2.2) Users provided with CFEs imagine this setting to be more helpful for others users, too (survey item 9).

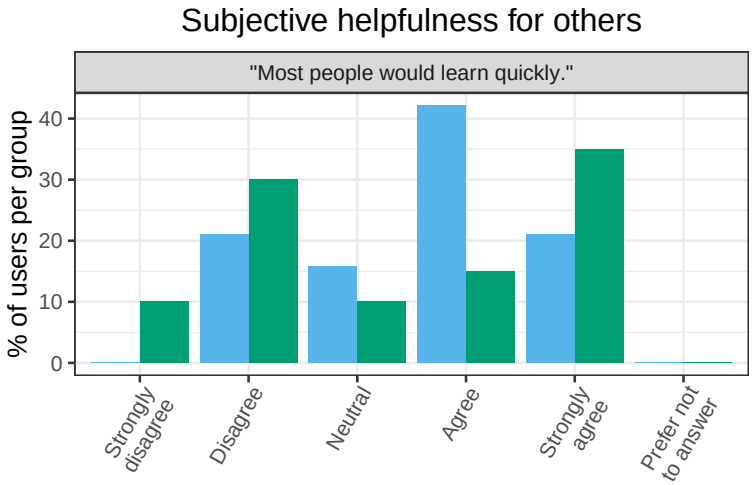
item 9: "I think most people would learn to work with the feedback on what choice would have led to a better result very quickly."

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

```
##     userId          group       itemNo  responseNo    checked
##  Length:234         C:114   Min.   :9   1:39       Min.   :0.0000
##  Class :character   E:120   1st Qu.:9   2:39       1st Qu.:0.0000
##  Mode  :character           Median :9   3:39       Median :0.0000
##                             Mean   :9   4:39       Mean   :0.1667
##                             3rd Qu.:9   5:39       3rd Qu.:0.0000
##                             Max.   :9   6:39       Max.   :1.0000
```

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

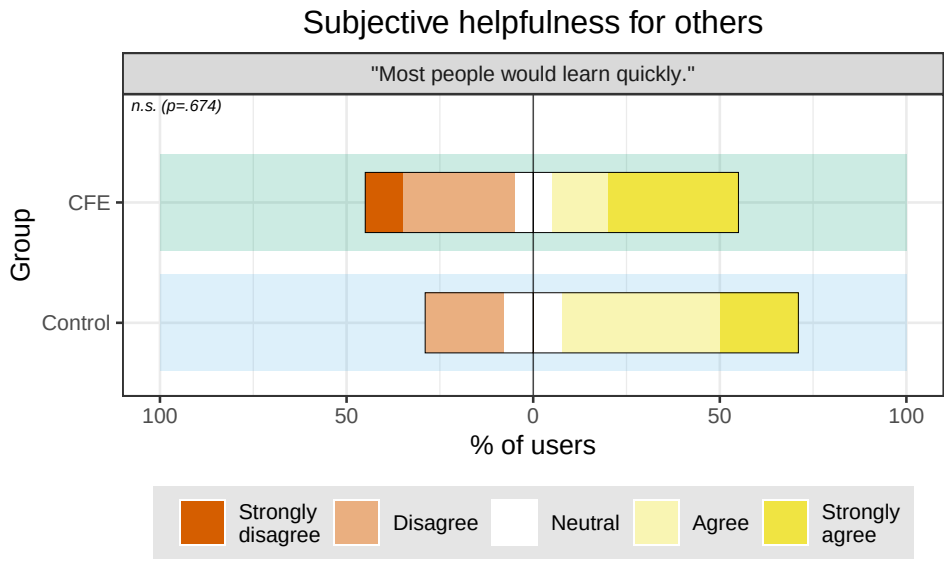




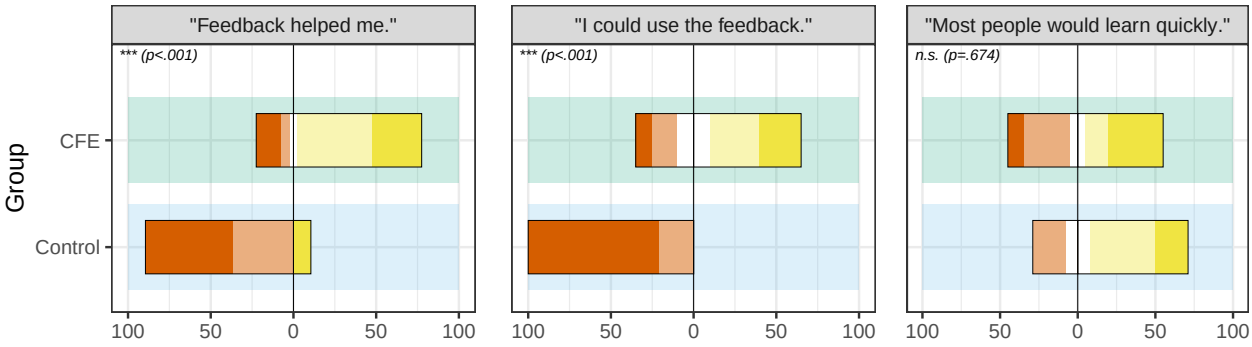
Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

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 explanation condition (M = 3.35, SEM = 0.3346247) and users in the control condition (M = 3.6315789, SEM = 0.2443577) using a Wilcoxon–Mann–Whitney U test. This showed: U=175, p=0.6737774, r = -0.0674091



H2) Final plot for publication



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:

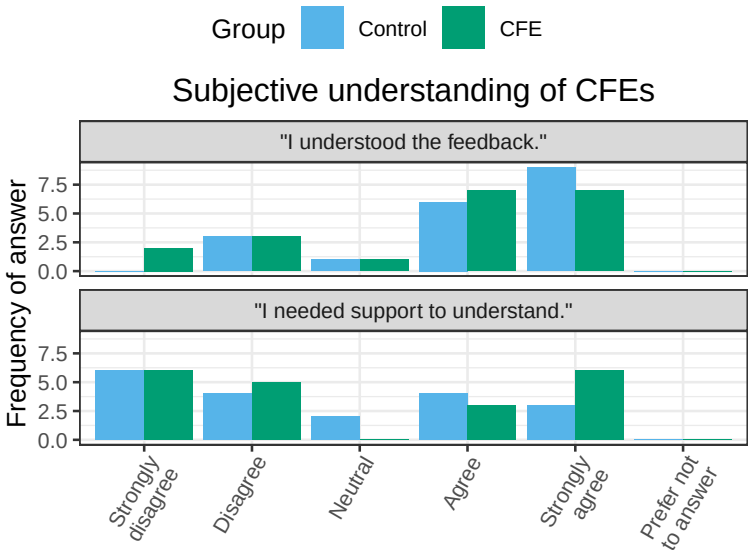
H3) We do not expect users in different conditions to differ in terms of how well they understood the feedback, or needing support for understanding (survey items 3, 4).

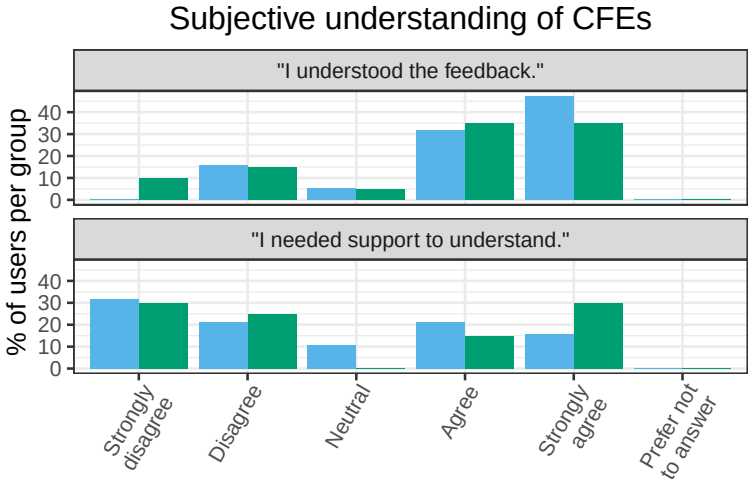
Item 3: "I understood the feedback on what choice would have led to a better result."

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:468         C:228   3:234   1:78       Min.   :0.0000
##  Class :character   E:240   4:234   2:78       1st Qu.:0.0000
##  Mode  :character                   3:78       Median :0.0000
##                                     4:78       Mean   :0.1667
##                                     5:78       3rd Qu.:0.0000
##                                     6:78       Max.   :1.0000
```

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



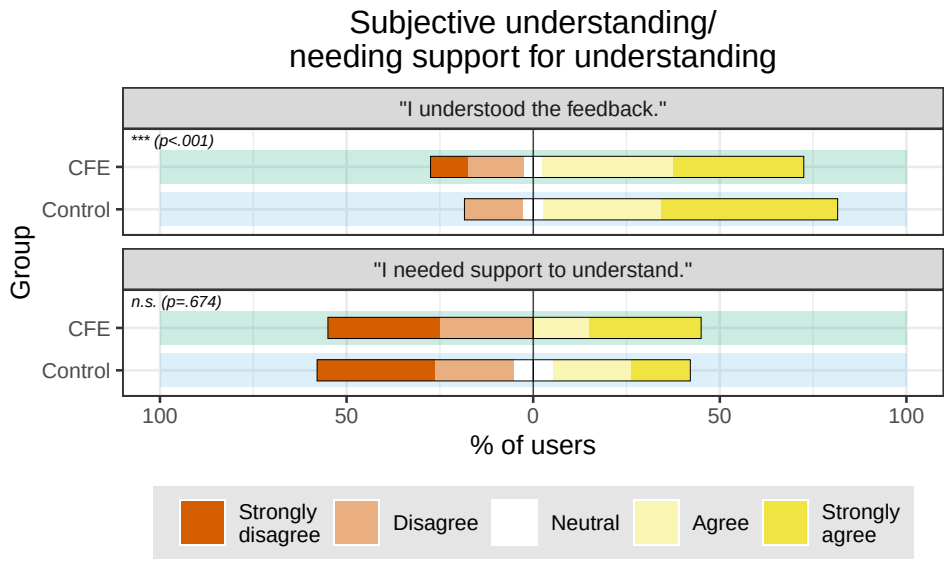


On to the statistical comparison: for Likert-scale, we want a non-parametric statistial test for ordinal data, that's the Wilcoxon–Mann–Whitney U test.

The analysis revealed:

- Is there a significant difference in terms of understanding of feedback between groups? We compared responses of users in explanation condition (M = 3.7, SEM = 0.3086473) and users in the control condition (M = 4.1052632, SEM = 0.2524122) using a Wilcoxon–Mann–Whitney U test. This showed: U=312.5, p=0.0003969, r = 0.5671965
- Is there a significant difference in terms of needing support to understand feedback?: We compared responses of users in explanation condition (M = 2.9, SEM = 0.3831998) and users in the control condition (M = 2.6842105, SEM = 0.3508772) using a Wilcoxon–Mann–Whitney U test. This showed: U=205, p=0.6744893, r = 0.067253

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



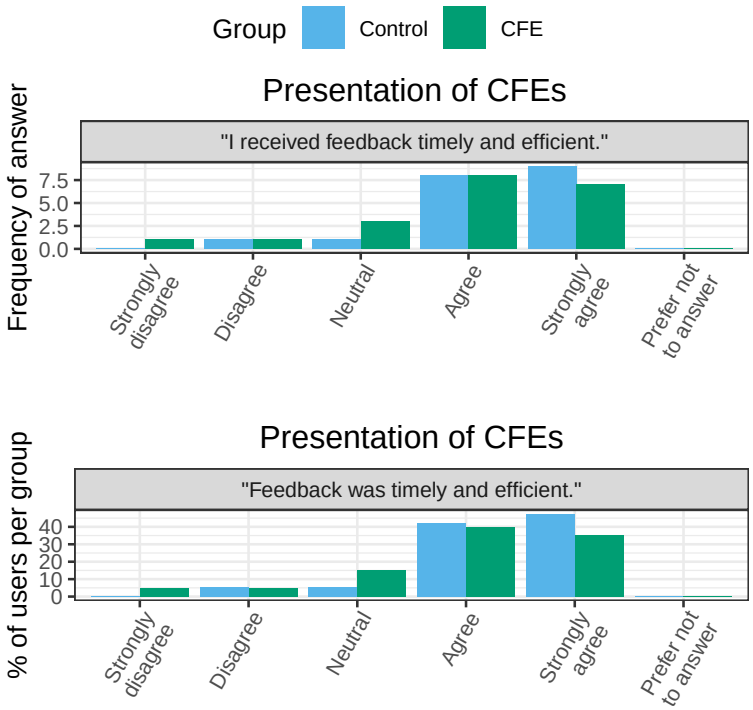
H4) Timing and efficacy of how CFEs were presented expected to be comparable

H4) We expect timing and efficacy of how CFEs were presented to be comparable, as it was literally the same (survey 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:234         C:114   Min.   :10   1:39       Min.   :0.0000
##  Class :character   E:120   1st Qu.:10   2:39       1st Qu.:0.0000
##  Mode  :character           Median :10   3:39       Median :0.0000
##                             Mean   :10   4:39       Mean   :0.1667
##                             3rd Qu.:10   5:39       3rd Qu.:0.0000
##                             Max.   :10   6:39       Max.   :1.0000
```

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

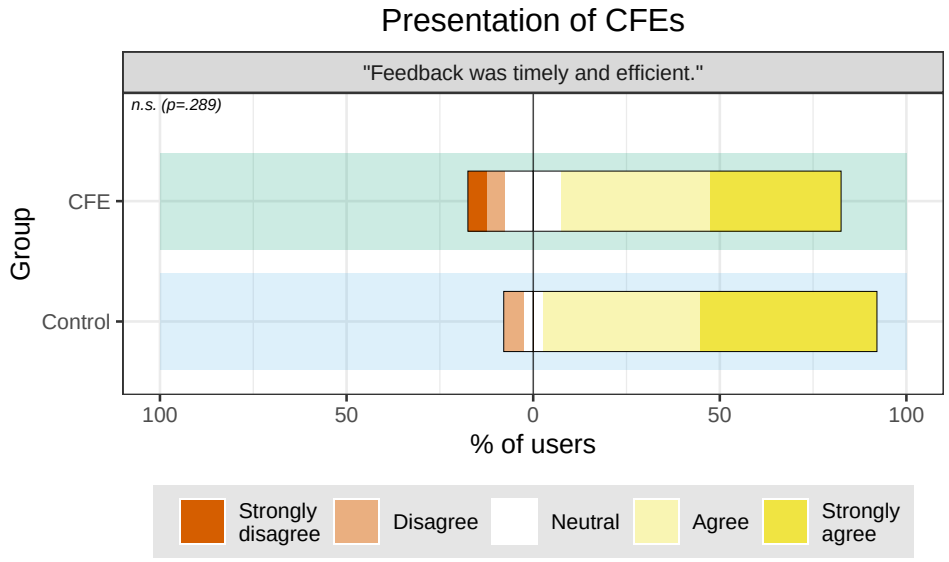


Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

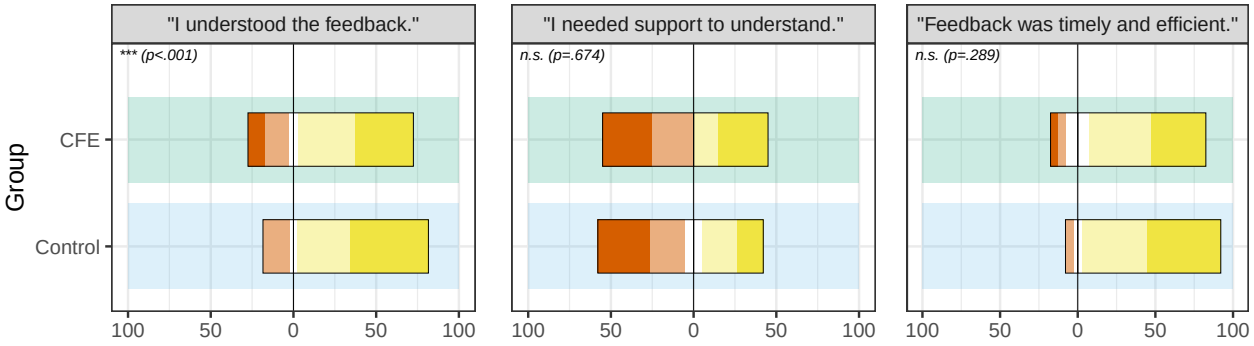
The analysis revealed:

- Is there a significant difference in terms of estimated usefulness for others between groups? We compared number of matches for users in explanation condition (M = 3.95, SEM = 0.2457534) and users in the control condition (M = 4.3157895, SEM = 0.1881369) using a Wilcoxon–Mann–Whitney U test. This showed: U=154.5, p=0.2892887, r = -0.1696851

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




```
## [1] "Summary plot for H3:"
```



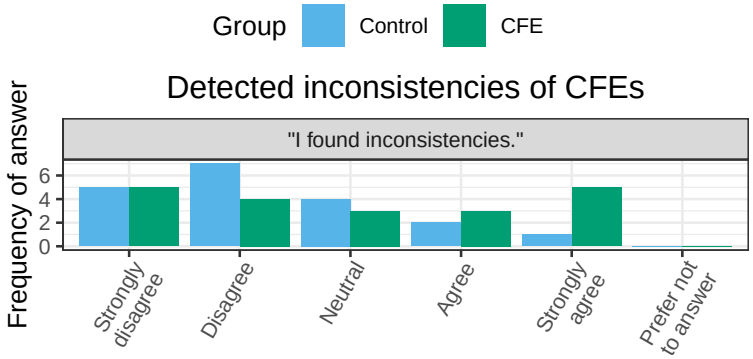
Final exploratory analysis

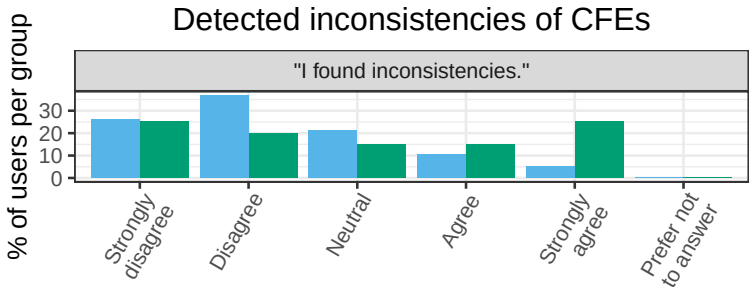
It is not clear whether users uncovered inconsistencies in the feedback. Let's ask them.

Item 8: "I found inconsistencies in the feedback presented."

```
##     userId           group       itemNo     responseNo    checked
##  Length:234         C:114   Min.   :8   1:39       Min.   :0.0000
##  Class :character   E:120   1st Qu.:8   2:39       1st Qu.:0.0000
##  Mode  :character           Median :8   3:39       Median :0.0000
##                             Mean   :8   4:39       Mean   :0.1667
##                             3rd Qu.:8   5:39       3rd Qu.:0.0000
##                             Max.   :8   6:39       Max.   :1.0000
```

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





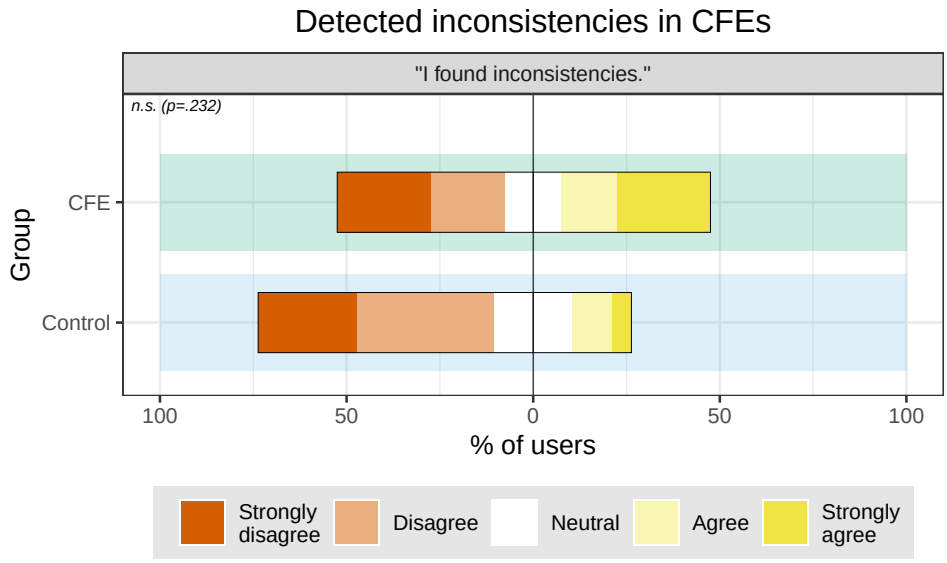
Check for significant differences between groups using the Wilcoxon–Mann–Whitney U test, as we have Likert-scale data.

The analysis revealed:

- Is there a significant difference in terms of found inconsistencies in the feedback provided? We compared answers of users in explanation condition (M = 2.95, SEM = 0.3515005) and users in the control

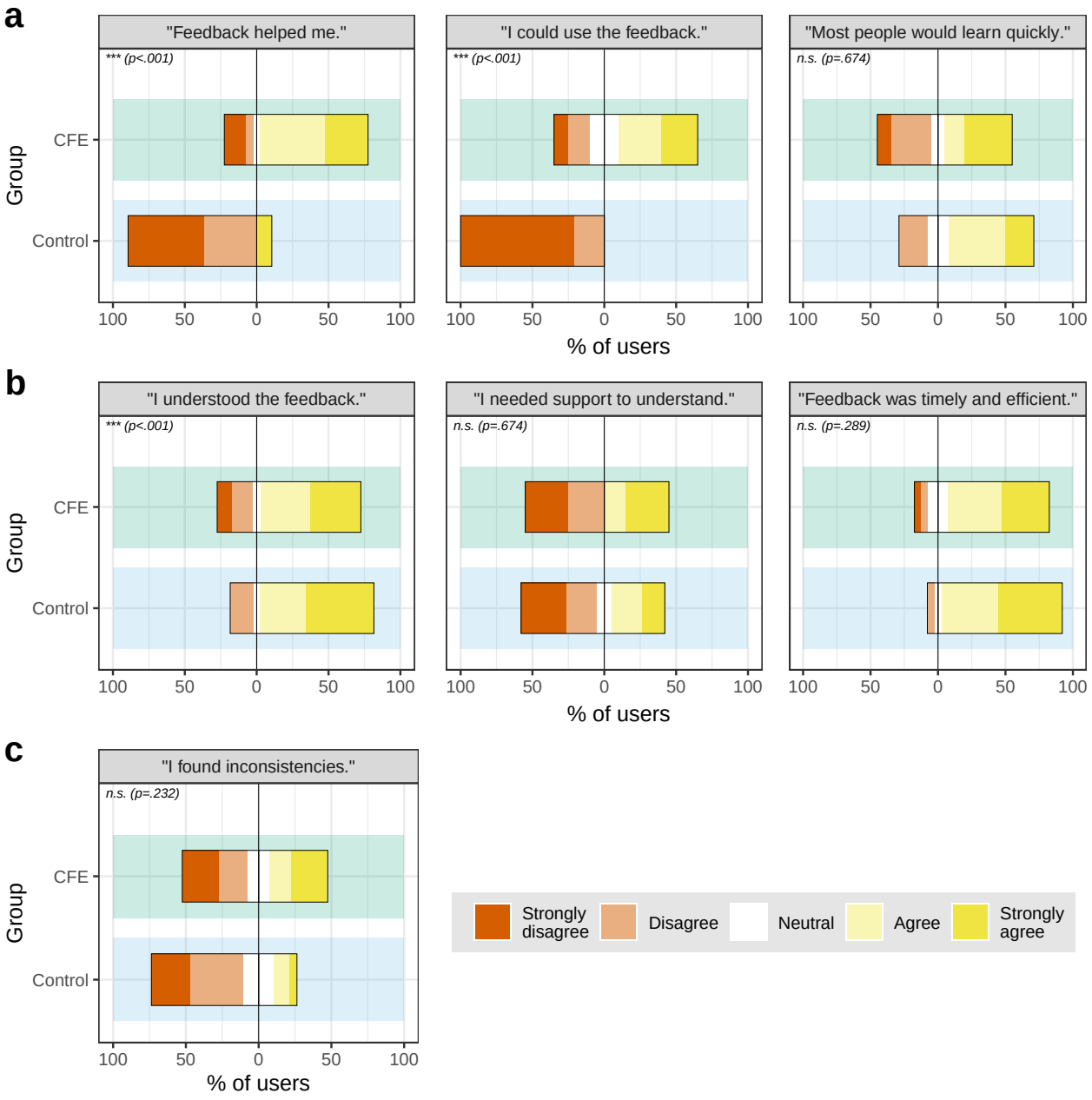
condition (M = 2.3157895, SEM = 0.2654868) using a Wilcoxon–Mann–Whitney U test. This showed: U=232, p=0.2315136, r = 0.1915883

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



Survey data: Final plot for publication

```
## [1] "Final summary plot for survey data:"
```



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.

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.